Virologisches Institut und
Abteilung Dermatologie, Klinik für Kleintiermedizin
der Vetsuisse-Fakultät Universität Zürich

Direktion Virologisches Institut: Prof. M. Ackermann
Leitung Abteilung Dermatologie: Prof. C. Favrot

Arbeit unter wissenschaftlicher Betreuung von Dr. Kurt Tobler

Viral and cellular transcription profiles in Equine Papillomavirus Type 2 positive squamous cell carcinomas

Inaugural-Dissertation

zur Erlangung der Doktorwürde der
Vetsuisse-Fakultät Universität Zürich

vorgelegt von

Anna Sophie Ramsauer

Tierärztin
von München, Deutschland

genehmigt auf Antrag von

Prof. Dr. Mathias Ackermann, Hauptreferent 1
Prof. Dr. Claude Favrot, Hauptreferent 2
Prof. Dr. Volker Thiel, Ko-Referent

2015

TABLE OF CONTENTS

Summary

Vetsuisse Faculty, University of Zurich 2015
Anna Sophie Ramsauer
Institute of Virology, email@vetvir.uzh.ch
and
Clinic for Small Animal Internal Medicine, Dermatology, sekretariat@kltmed.ch

Viral and cellular transcription profiles in Equine Papillomavirus Type 2 positive squamous cell carcinomas

Among the different equine papillomaviruses (EcPV), EcPV type 2 seems to have the most important clinical impact on horses, since it appears to be a causal factor for the development of penile squamous cell carcinomas (SCC). Unfortunately, the pathomechanisms associated with this cancer transformation are not known, yet.
To address this issue, penile tissue samples were collected from five horses with EcPV2-positive SCC as well as from three healthy EcPV2-negative horses. Thereof transcriptome analysis of the viral and the host genome was performed.
Indeed, EcPV2 was transcriptionally active in all the SCC samples, whereas EcPV2 mRNA was not detected in any control tissues. While only few reads mapped to the structural viral genes, up to some thousands reads mapped to the non-structural early (E) genes, in particular to E2, E4, E6 and E7. Within these reads a distinct pattern of splicing events was observed, while most abundant splice junctions were E1^E4 and E1^E2. Host genes, differentially expressed between SCC and control samples, were most abundantly related to cell cycle, RNA processing and focal adhesion. A comparison of our findings with published results suggested that Matrix Metalloproteinase 1 (MMP1) and Interleukin 8 (IL8) may represent potential marker genes for the development of SCCs in EcPV2 associated lesions.
Consequently, further research on EcPV2-related pathomechanisms may focus on the viral genes E2, E4, E6 and E7 and the cellular genes MMP1 and IL8.

Keywords: horse, papillomavirus, EcPV2, penile carcinoma, transcriptome

Zusammenfassung

Vetsuisse-Fakultät Universität Zürich 2015
Anna Sophie Ramsauer
Virologisches Institut, email@vetvir.uzh.ch
und
Klinik für Kleintiermedizin, Dermatologie, sekretariat@kltmed.ch

Virales und zelluläres Transkriptionsprofil von Equines-Papillomavirus-Typ-2-positiven Plattenepithelkarzinomen

Unter den verschiedenen equinen Papillomaviren (EcPV) scheint EcPV Typ 2 die größte klinische Relevanz beim Pferd zu haben, da es offenbar ein entscheidender Faktor für die Entwicklung von Plattenepithelkarzinomen (PEK) am Penis ist. Allerdings sind die genauen Pathomechanismen dieser Krebstransformation leider noch nicht bekannt.

Daher wurden Gewebeproben vom Penis von fünf Pferden mit EcPV2-positiven PEKs, sowie von drei gesunden EcPV2-negativen Pferden gesammelt. Davon wurden Transkriptom-Analysen des Virus- und des Wirtsgenom durchgeführt.

Transkripte von EcPV2 konnten in allen PEK Proben nachgewiesen werden, jedoch nicht in den Kontroll-Proben. Während die viralen Strukturgene nur spärlich von der Sequenzinformation abgedeckt wurden, ließen sich mehrere tausend Sequenzstücke den Nichtstrukturgenen (E), vor allem E2, E4, E6 und E7, zuordnen. Dort konnten außerdem Spleißverbindungen detektiert werden, wobei E1^E4 und E1^E2 die häufigsten waren. Die Wirtsgene, die zwischen PEKs und Kontrollproben unterschiedlich exprimiert waren, stehen vor allem in Zusammenhang mit Zellzyklus, RNA-Prozessierung und fokaler Adhäsion. Vergleiche mit anderen Studien zeigten, dass Matrix Metalloproteinase 1 (MMP1) und Interleukin 8 (IL8) als Markergene für die Entstehung von PEKs in EcPV2 assoziierten Veränderungen dienen könnten.

Daher möge sich die weitere Erforschung der EcPV2 assoziierten Pathomechanismen auf die viralen Gene E2, E4, E6 und E7 und die zellulären Gene MMP1 und IL8 fokussieren.

Stichworte: Pferd, Papillomavirus, EcPV2, Peniskarzinom, Transkriptom

Abbreviations

BPV	bovine papillomavirus
CDK	cyclin dependent protein kinase
CIN	cervical intraepithelial neoplasia
CXCL8	chemokine (CXC motif) ligand 8
DE	differently expressed
E	early gene (papillomavirus)
EcPV	equine papillomavirus
fc	fold change
FDR	false discovery rate
GEO	Gene Expression Omnibus
GO term BP	Gene Ontology term biological processes
HPV	human papillomavirus
IL8	Interleukin 8
KEGG	Kyoto Encyclopedia of Genes and Genomes
L	late gene (papillomavirus)
LCR	long control region
MMP1	Matrix Metalloproteinase 1
NCBI	National Center for Biotechnology Information
ORF	open reading frames
PEK	Plattenepithelkarzinom
PV	papillomavirus
RIN	RNA integrity number
SCC	squamous cell carcinoma
TMM	trimmed means of M values

1. Introduction

1.1. Papillomavirus

Papillomaviruses (PV) are non-enveloped highly species-specific epitheliotropic DNA viruses (Howley PM, Lowy DR et al. 1990). After infection of keratinocytes in the basal layer of stratified squamous epithelia, they replicate in the nucleus in a differentiation-dependent manner (Zheng, Baker 2006). They can induce benign lesions of the skin and mucous membranes, like warts and condylomas in many species. Some PVs are also involved in the development of malignant epithelial lesions like squamous cell carcinomas (SCCs) (Howley PM, Lowy DR et al. 1990). The cancer-associated human papillomaviruses (HPV) are classified as high-risk HPV types, while the other HPVs are divided into intermediate or low risk types depending on the frequency with which they are found in cancers (Doorbar 2006).

The PV particles are about 55 nm in diameter and consist of a single molecule of double-stranded circular DNA, surrounded by a capsid. The viral genome, which is approximately 8000 base pairs in size, contains up to ten open reading frames (ORFs) and a long control region (LCR). The ORFs are classified as early genes (E), which encode for viral proteins involved in the regulation of replication and synthesis of viral DNA, and late genes (L), which encode for the structural capsid proteins (Howley PM, Lowy DR et al. 1990). Usually the viral gene expression leads to the expression of up to six non-structural viral regulatory proteins (E1, E2, E4, E5, E6 and E7) in undifferentiated or intermediately differentiated keratinocytes and two structural viral capsid proteins (L1 and L2) in keratinocytes undergoing terminal differentiation (Zheng, Baker 2006).

The viral gene expression is tightly regulated at transcriptional and post-transcriptional levels depending on the cell differentiation in the infected cells (Zheng, Baker 2006). In general, transcription of PV starts at either of two main promoter and end at either of two poly(A) sites revealing two main polycistronic pre-mRNA species. Through extensive use of alternative splicing PVs have the ability to produce a variety of products from the same genomic sequences from their compact genomes and to individually regulate expression of each gene during the viral life cycle (Schwartz 2013).

After infection of keratinocytes in the basal layer, the virus expresses E1 and E2 proteins, which are needed for replicating and maintaining the viral DNA and the regulation of early transcription. The E4 protein continues to be expressed in the terminally differentiated keratinocytes and is associated with cytokeratin filament collapse. E5, E6, and E7 are viral oncogenes, however not all PVs encode for an E5 protein. They induce cell immortalization and transformation. In particular the viral oncoproteins E6 and E7 stimulate cell growth by inactivating the cellular tumour suppressor proteins p53 and pRb. The late genes L1 and L2 are needed to assemble the capsids in the upper layers (Doorbar 2005; Zheng, Baker 2006).

The lifecycle of PV depends on epithelial differentiation. The viral genomes are maintained as episomes in the basal layer and the viral gene expression is closely regulated by the migration of infected cells towards the epithelial surface. However, in some high-risk PVs the viral gene expression becomes deregulated and the normal life cycle of the virus cannot be completed (abortive infection). The dysregulation might be correlated with loss of E2 gene due to integration of the genome into the host genome. Such abortive infections can predispose to cancer, as with other DNA tumour viruses. The cancer develops in individuals who fail to resolve their infection and retain oncogene expression for years or decades. Though, in most individuals, immune regression leads to clearance or maintenance in a latent or asymptomatic state in the basal cells (Doorbar 2005, 2006).

1.2. Papillomavirus in horses

Seven different equine papillomaviruses (EcPV1-7) and three bovine papillomaviruses (BPV1, BPV2 and BPV13) are described to infect horses (Lange et al. 2013b; Scott, Miller 2011; Lange et al. 2011; Ghim et al. 2004; Lunardi et al. 2013). The bovine papillomaviruses are associated with the equine sarcoid and induce fibropapillomas (Scott, Miller 2011; Lunardi et al. 2013; Howley PM, Lowy DR et al. 1990). EcPV1 causes the so called classical equine viral papillomas, which occur mainly on the muzzle or lips and are typically found in young horses (Scott, Miller 2011). EcPV2 was detected in penile papillomas and SCC, while EcPV3 seems to be associated with aural papillomas and plaques (Lange et al. 2011). Yet, there is not much known about the other EcPVs, though they were also detected in genital plaques (EcPV4), aural plaques (EcPV5, EcPV6) or penile masses (EcPV7) (Lange et al. 2013b). Within the different equine papillomaviruses, EcPV2 seems to have the

most important clinical impact on horses, since it appears to be a causal factor for the development of penile squamous cell carcinomas (Bogaert et al. 2012; Knight et al. 2011; Lange et al. 2011; Lange et al. 2013a; Scase et al. 2010; Sykora et al. 2012).

1.3. Equine penile squamous cell carcinoma

SCC is a malignant neoplasm arising from keratinocytes. It occurs preferentially at mucocutaneous junctions, most frequently at the external genitalia and the ocular region in horses (Scott, Miller 2011). Penile and preputial tumours are not uncommon in horses. Several tumour types have been described at the external genitalia of horses, but SCC represents the most frequent genital neoplasm. Penile SCCs are mainly detected on the glans penis (Van Den Top, J G B et al. 2008). Early clinical, stages present as grey small papules while advanced cases may develop more proliferative or ulcerated lesions (Lange et al. 2013a). Genital SCCs pose a significant welfare problem to the affected horses, because aggressive and often repeated surgical therapy is required to manage the lesions. In most cases phallectomy is the treatment of choice (Mair et al. 2000). The prevalence of SCC is significantly higher in geldings, light pigmented breeds and also increases with the age (Scott, Miller 2011). The aetiology of penile SCC in horses has not completely been investigated so far, but there is strong evidence that EcPV2 seems to be involved in the development of those lesions (Scase et al. 2010; Lange et al. 2013a; Bogaert et al. 2012; Knight et al. 2011; Sykora et al. 2012; Lange et al. 2011).

1.4. EcPV2

EcPV2 was first described only a few years ago. Two closely related sequence variants of this virus were published in 2010 and 2011 (Scase et al. 2010; Lange et al. 2011). The EcPV2 genome consists of 7803 basepairs and contains 7 ORFs, a large and a small non-coding region (Scase et al. 2010; Lange et al. 2011). The ORFs encode for five early PV viral genes E6, E7, E1, E2 and E4 and two late structural viral genes L2 and L1 (Scase et al. 2010; Lange et al. 2011). Since the first description of the virus, EcPV2 DNA was consistently detected in equine penile papillomas and in situ and invasive SCCs (Bogaert et al. 2012; Knight et al. 2011; Lange et al. 2011; Lange et al. 2013a; Scase et al. 2010; Sykora et al. 2012). However, EcPV2 DNA was also detected in the mucosa of healthy horses. The prevalence for EcPV2 differs depending on the different studies between 2.6% and

18% (Fischer et al. 2014; Sykora et al. 2012; Bogaert et al. 2012; Knight et al. 2011; Knight et al. 2013). In Switzerland more than one third of the healthy horses have antibodies against EcPV2. Possibly EcPV2 is a prevalent virus circulating among the equine population and other cofactors, similar to the situation during HPV infections, could be involved in the development of SCCs (Fischer et al. 2014). Nevertheless EcPV2 is significantly more frequently detected in penile SCCs than in non-cancerous tissues and also shows higher viral loads in SCCs (Knight et al. 2013). Viral mRNA, which suggests an active pathogenic role of the virus or the viral proteins in the formation of SCCs could be shown in these lesions (Sykora et al. 2012). Yet, there is not much known, about the relationship between tumour characteristics and prognosis (Van den Top, J. G. B. et al. 2014). In human penile carcinomas the presence of high-risk human papillomavirus DNA predicts favourable outcome in survival (Lont et al. 2006). Also the nuclear proteins tumour protein p53 and proliferation marker Ki67 seem to be correlated to nodal metastasis and are suggested as prognostic markers in human penile cancer (Muneer et al. 2009; Protzel et al. 2007). However, neither p53 and Ki67 nor the presence of EcPV2 or the expression of the viral genes (E2, E6 and E1) seem to be good prognosticators for equine penile SCCs (Van den Top, J. G. B. et al. 2014). Until now there is a lack of knowledge about good prognosticators and the mechanisms driving this disease associated with this prevalent virus. Therefore it might be important to unravel the pathomechanisms associated with the cancer transformation and to identify genes which could be useful as biomarkers for the development and prognosis of EcPV2 associated SCC.

1.5. RNA-Seq

RNA-Seq is a recently developed approach, which allows transcriptome profiling of genomes at a high resolution (Nagalakshmi et al. 2010; Wang et al. 2009). Complementary DNAs (cDNAs) generated from the RNA pool of interest are directly sequenced using deep sequencing technologies. By aligning the obtained millions of reads to a reference genome a transcriptome map can be constructed. Furthermore, counting the reads matching to defined gene locations reveals their transcription levels (Nagalakshmi et al. 2010).

1.6. Aim of the study

The limited knowledge of EcPV2 and its association to develop SCC, urge to intensify studies of this infectious agent. In particular, the changes in gene expression of affected tissue might give new perspectives into the pathological mechanism of the associated disease. EcPV2 transcription during infection, as well as, differences between host gene expression in healthy horses and horses with EcPV2 positive penile SCCs can be determined by RNA-Seq. Information about EcPV2 on the transcriptional level might help to further characterize this virus, while the categorization of affected biological processes and pathways in the host will open a new perspective of the molecular events leading to SCC. Furthermore, potential marker genes for the future development of EcPV2 positve SCCs in the affected penis of horses will be defined. The underlying hypothesis is that the gene expression profile of EcPV2 infected tissue is altered as compared to healthy skin cells. The overall gene expression pattern in EcPV2 affected SCC will give clues about the mechanisms driving the disease. This study will provide a basis to study EcPV2 associated cell changes in more detail and might be useful for the future development of diagnostic and prognostic tools in EcPV2 associated SCCs.

2. Material and methods

2.1. Sample material

Penile skin biopsy samples were collected from five horses with penile squamous cell carcinoma, as well as from three horses, which were apparently free of any skin lesions. Sampled affected horses were presented for diagnosis and treatment of SCC in different horse-clinics in Switzerland or Germany. The samples were either taken from diagnostic biopsies or during surgery after excision of the tumour. The clinical diagnosis for squamous cell carcinoma was confirmed by histopathological examination and the presence of EcPV2 DNA was evaluated by PCR assay. The biopsy samples from unaffected horses were collected from recently euthanized or slaughtered horses from a Swiss butcher or from horse-clinics in Switzerland or Germany. All collected tissue material was immediately immersed in RNA-later (SIGMA-ALDRICH, Buchs, CH) and incubated for 24 hours at 4°C. Then RNA-later was removed and samples were stored at -70°C until further processing.

2.2. DNA Extraction and PCR

DNA was extracted from an approximately two millimetre sized piece of the biopsy using QIAamp DNA Minikit (QIAGEN, Germantown, MD) according to the manufacture`s protocol. Successful genomic DNA extraction was confirmed by a PCR assay designed to amplify equine GAPDH DNA with the primer combination ecGAPDH forward primer (5'-GAG CTG AAT GGG AAG CTC AC-3') and reverse primer (5'-CTG AGG GCC TTT CTC CTT CT-3') (Lange et al. 2013a). For detection of EcPV2-DNA, a PCR assay was performed using an EcPV2 specific primer set. This assay amplified a 395 base pair fragment of the E6 gene (position 41-435) from the genomic sequence of EcPV2 (Gen Bank accession number EU503122), using EcPV2-E6 forward primer (E6-41f: 5′-GCTCTCTTCTGGTTACTTTGG-3′) and reverse primer (E6-435r: 5′-CTTGCAGTGTACACGTTTCTG-3′) (Scase et al. 2010). PCR was performed with a reaction mix containing 8 µl water, 2 µl of each forward and reverse primer (10µM each), 1µl extracted DNA as template and 12 µl REDTaq ReadyMIX (SIGMA-ALDRICH, Buchs, CH) in a total volume of 25 µl. The cycling program for both PCR assays started with a denaturation step of 3 min at 94°C, followed by 40 cycles of 30 sec at 94°C, 30 sec at 55°C and 30 sec at 72°C. PCR products were separated by electrophoresis through a 1% agarose gel and visualized

by ethidium bromide staining. The PCR amplimers were excised from the agarose gel and purified using Zymoclean Gel DNA Recovery Kit (ZYMO RESEARCH, Irvine, CA) according to the manufacture´s protocol. Nucleotide sequences were determined (Microsynth, Balgach, CH) using an ABI 377 sequencer (Applied Biosystems/Invitrogen, Basel, CH) and compared to the published reference sequences of EcPV2 with the NCBI Basic Local Alignment Search Tool ("BLAST") (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) (Altschul et al. 1990).

2.3. RNA Extraction

An approximately 2 mm sized piece of the biopsy was cut into small pieces and placed in a ZR BashingBead Lysis Tube (ZYMO RESEARCH, Irvine, CA) containing 500 µl TRIzol Reagent (life technologies, Zug, CH). The sample was homogenized by vortexing for at least ten minutes. Afterwards the RNA was extracted using Direct-zol RNA MiniPrep (ZYMO RESEARCH, Irvine, CA) according to the manufacture´s protocol including an in-column DNase I digestion. The quantity and quality of the extracted RNA was assessed by the use of a Bioanalyzer 2100 (Agilent, Waldbronn, Germany). Samples with a RNA Integrity Number (RIN) of at least 7.2 were included in RNA-Seq analyses. For all extracts 400ng were used for further processing by RNA-Seq.

2.4. Illumina RNA Sequencing

2.4.1. Library preparation

The quantity and quality of the isolated RNA was determined with the Qubit® (1.0) Fluorometer (Life Technologies, California, USA) and the Bioanalyzer 2100 (Agilent, Waldbronn, Germany). The SureSelect SS RNA Library Prep (Agilent, Waldbronn, Germany) was used in the succeeding steps. Briefly, total RNA samples (1 µg) were ribosome depleted and then reverse-transcribed into double-stranded cDNA with Actinomycin added during first-strand synthesis. The cDNA samples was fragmented, end-repaired and polyadenylated before ligation of TruSeq adapters. The adapters contain the index for multiplexing. Fragments containing TruSeq adapters on both ends were selectively enriched with PCR. The quality and quantity of the enriched libraries were validated using Qubit® (1.0) Fluorometer and the Bioanalyzer 2100 (Agilent, Waldbronn, Germany). The product is a smear with an average fragment

size of approximately 360 base pairs. The libraries were normalized to 10 nM in Tris-Cl 10 mM, pH 8.5 with 0.1% Tween 20.

2.4.2. Cluster Generation and Sequencing

The TruSeq SR Cluster Kit v3-cBot-HS (Illumina, Inc, California, USA) was used for cluster generation using 8 pM of pooled normalized libraries on the cBOT. Sequencing was performed on the Illumina HiSeq 2000 in single read mode with read length of 100 base pairs using the TruSeq SBS Kit v3-HS (Illumina, Inc, California, USA)

2.5. Sequencing data analysis

After 101 cycles of single-end sequencing, the processing of fluorescent images into sequences, base-calling and quality value calculations were performed using the Illumina data processing pipeline (version 1.8.2).

The raw reads were first cleaned by removing adapter sequences, trimming low quality ends, and filtering reads with low quality (phred quality <20). Sequence alignment of the resulting high-quality reads to the Equine (build: EquCab2) and viral genome (EcPV2_GB: EU503122.1) and quantification of gene level expression was carried out using “RSEM” (Version 1.2.5) (Li, Dewey 2011). To detect differentially expressed genes we applied count based negative binomial model implemented in the software package “edgeR” (R version: 3.0.2, edgeR version: 3.4.2) (Robinson et al. 2010), in which the normalization factor was calculated by trimmed mean of M values (TMM) method (Robinson, Oshlack 2010). The gene-wise dispersions were estimated by conditional maximum likelihood and an empirical Bayes procedure was used to shrink the dispersions towards a consensus value. The differential expression was assessed using an exact test adapted for over-dispersed data. Genes showing altered expression with adjusted (Benjamini and Hochberg method) p-value < 0.05 were considered differentially expressed.

2.6. Analysis viral transcriptome

To characterize the viruses and their transcription within the EcPV2 positive horses, in particular, the quantity of the reads, the sequence variations and the splicing of the viral mRNA were analysed and represented by Integrative Genomics Viewer “IGV” (http://www.broadinstitute.org/igv/) (Robinson et al. 2011; Thorvaldsdóttir et al. 2013).

2.7. Analysis horse transcriptome

The differently expressed genes in EcPV2 positive horses with SCC compared to healthy horses with p-values below 0.01 and fold changes above 2 ($\triangleq |log2\ ratio| > 1$) were represented in an histogram and used in an unsupervised cluster analysis by "R" (http://www.R-project.org/) (R Core Team; R Development Core Team 2013). Fold changes (fc) were defined as fc = $2^{|log_2 ratio|}$.

2.7.1. Pathway and Biological Processes analysis

Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway and Gene Ontology (GO) term biological processes (BP) analysis was done by three different programs: "GSEA" (http://www.broadinstitute.org/gsea) (Mootha et al. 2003; Subramanian et al. 2005), "STRING" (http://string-db.org) (Franceschini et al. 2013; Szklarczyk et al. 2011) and "DAVID" (http://david.abcc.ncifcrf.gov) (Huang et al. 2009b, 2009a). A list of genes with a p-value < 0.01 and fc > 2 was used for pathway analysis by "DAVID" and "STRING". At the "GSEA" analysis all genes were included depending on the different method the program is using. The pathways and terms detected within all the three different programs were decided to be the most essential ones. Selected pathways were represented using R Bioconductor package "PATHVIEW" (http://www.bioconductor.org/packages/release/bioc/html/pathview.html) (Luo, Brouwer 2013). An Enrichment Map of the "GSEA" GO terms BP analysis was constructed using "CYTOSCAPE" (Enrichment Map Plugin) (http://www.cytoscape.org) (Saito et al. 2012; Cline et al. 2007).

2.7.2. Detection of potential Marker genes

To detect potential marker genes a gene list of the TOP 100 up regulated genes with the lowest p-values were compared via "EXPRESSIONBLAST" (http://www.expression.cs.cmu.edu) (Zinman et al. 2013) against processed data sets published on Gene Expression Omnibus (GEO) on the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/geo/). Once, analysis was done versus the entire database and once versus datasets with keyword "carcinoma" to look especially for carcinoma studies.

3. Results

3.1. Horses

Tissue samples from eight horses were included for RNA-Seq analysis. One group consisted of three healthy EcPV2-negative horses and the other group of five horses with EcPV2-positive SCCs. More details about these horses are listed in Table 1. The mean and median ages of the two groups (mean SCC 18.2 and mean healthy 20 years) as well as this of eight horses in total (mean 18.8 years) are similar. All horses were geldings. Efforts were made to match the breeds and origins represented in both healthy and affected group (see Table 1). Every penile lesion was a proliferative SCCs and the clinical diagnosis was confirmed by histopathological examination.

3.2. DNA extraction and PCR

To check for the presence of viral DNA, from every tissue sample, total DNA was extracted and checked for GAPDH-DNA and EcPV2-DNA by PCR. GAPDH-DNA could be amplified in all samples, so the successful DNA extraction was confirmed. As expected, EcPV2-DNA was detected by using an EcPV2_E6 specific primer pair in every sample from penile SCC, but not in the samples from healthy horses. (Table 1) The sequences of the PCR amplimers were determined. The sequencing results showed a similarity of minimum 98% compared to the published EcPV2 sequences by "BLAST" analysis (Status 2014) (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi.) (Altschul et al. 1990).

Table 1: Horse samples used in this study

Sample	Age	Origin	Bread	Clinics	PCR
1	19	Switzerland	Arab/Welsh Pony	healthy	negative
2	21	Germany	Quarter Mix	healthy	negative
3	20	Switzerland	Warmblood	healthy	negative
4	16	Germany	Haflinger	SCC	EcPV2
5	14	Germany	New Forest Pony	SCC	EcPV2
6	21	Switzerland	Warmblood	SCC	EcPV2
7	18	Germany	Haflinger	SCC	EcPV2
8	22	Germany	Welsh Pony	SCC	EcPV2

3.3. RNA-Seq

3.3.1. Viral transcriptome

RNA-Seq analysis was employed to characterize the viruses and their transcription within the EcPV2 positive horses. In particular, the sequence variations and the splicing of the viral mRNA were analysed.

3.3.1.1. Viral reads

To check for the expression of the individual genes, the RNA-Seq reads were mapped to the viral genome (EcPV2_GB: EU503122.1.) Hits were detected for all samples of EcPV2 positive horses with SCC, but not for samples of healthy horses. The read counts of the five positive samples mapped to the EcPV2 genome are shown in Figure 1. The gene expression profile of EcPV2 was similar in all samples. In ORFs of the late genes only a few reads were noticed, while up to some thousands reads in ORFs of the early genes - in particular E6, E7, E4 and E2 - were detected.

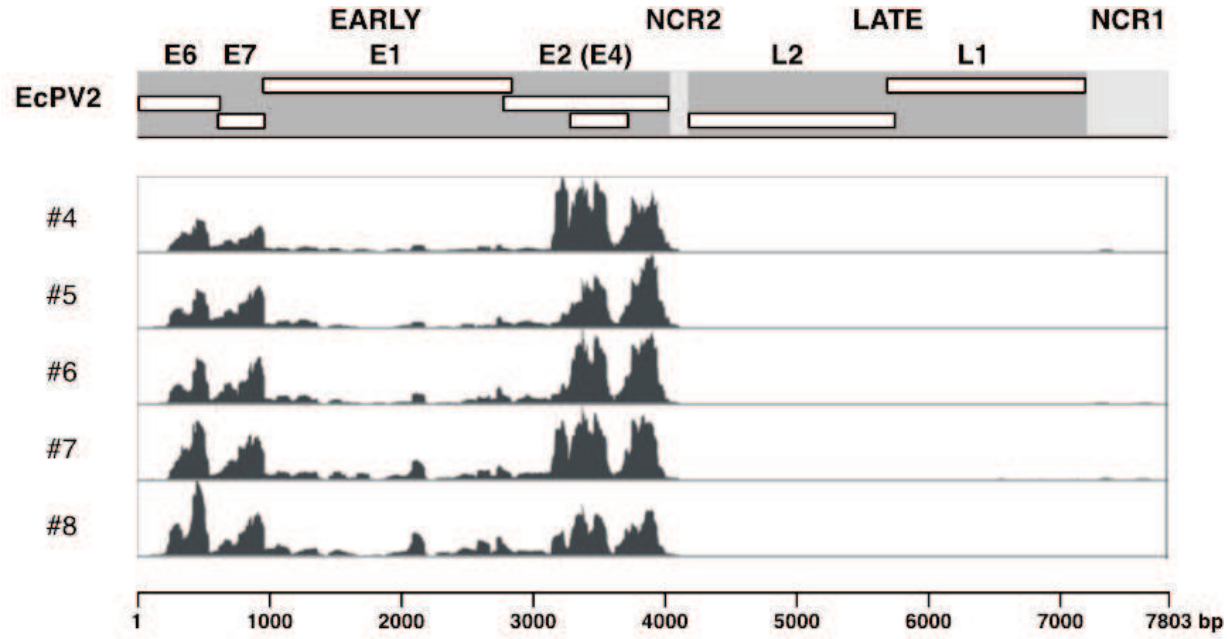
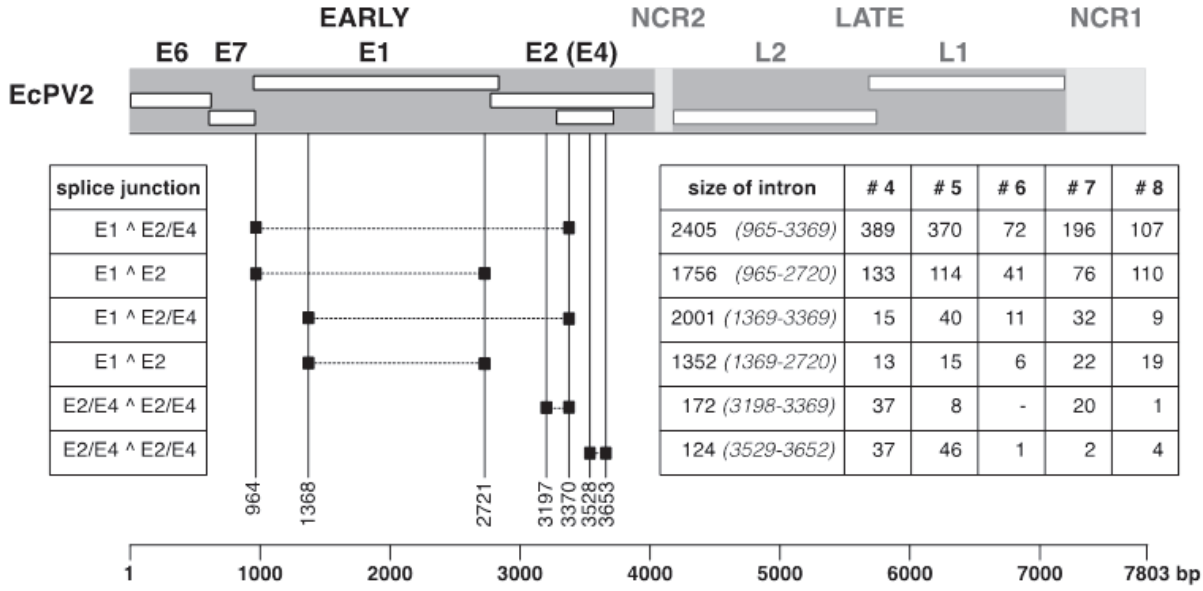


Figure 1: Comparison of the quantity of reads which mapped to the EcPV2 genome. *The RNA-Seq read counts of the five EcPV2 positive samples mapped to the EcPV2 genome below the schematic view of the EcPV2 genome. (scale of reads #4: 0-2719; #5: 0-1414; #6: 0-993; #7: 0-370, #8: 0-733). In ORFs of L genes only a few reads were noticed, while up to some thousands reads in ORFs of E genes were detected.*

3.3.1.2. Splice junctions EcPV2

Upon the analysis of the viral reads we recognized that some reads could be mapped onto exon junctions. Based on these reads, the corresponding splice junctions could be evaluated. The six most frequent EcPV2 splice junctions found in the reads from the five positive horses and their associated read counts are shown in Figure 2. Most common were junctions between the E1^E2, E1^E2/E4 and E2/E4^E2/E4 region. Most prevalent were donors at 964 and 1368 and acceptors at 3370 and 2721. There were no intron-spanning reads detected in the late region due to the low sequence coverage in this region.



splice junction
E1 ^ E2/E4
E1 ^ E2
E1 ^ E2/E4
E1 ^ E2
E2/E4 ^ E2/E4
E2/E4 ^ E2/E4

size of intron	# 4	# 5	# 6	# 7	# 8
2405 *(965-3369)*	389	370	72	196	107
1756 *(965-2720)*	133	114	41	76	110
2001 *(1369-3369)*	15	40	11	32	9
1352 *(1369-2720)*	13	15	6	22	19
172 *(3198-3369)*	37	8	-	20	1
124 *(3529-3652)*	37	46	1	2	4

Figure 2: Splice-junctions within the EcPV2 reads.
The six most frequent EcPV2 splice junctions found in the reads from the five positive horses are shown below the schematic view of the EcPV2 genome. The numbering of the exon-junctions corresponds to the nucleotide numbering of EcPV2_ZH strain (HM461973). The table presents the not standardized read counts over introns in the particular samples.

3.3.1.3. Sequence variations in the different EcPV2 genomes

To characterize the EcPV2 variants in our samples, the sequences revealed from the RNA-Seq experiment were compared to the published EcPV2 variants. Thirty-seven positions different to the EcPV2_GB variant (EU503122.1) were detected in the five samples. These variations were also compared to the EcPV2_ZH (HM461973) variant. The resulting sequence variations are shown in Figure 2. There were no variations detected in the late region because of the low sequencing coverage

(Figure 1). The result of this analysis proposed that sample #4 was closer related to EcPV2_GB and the other four samples might have formed yet not recognized EcPV2 variants. In Detail there were four sequence variations in the E6 ORF, three in the E7 ORF, 18 in the E1 ORF, twelve in the E2 ORF and none in the E4 ORF. Thereof 23 were silent, while 14 were non-synonymous point mutations.

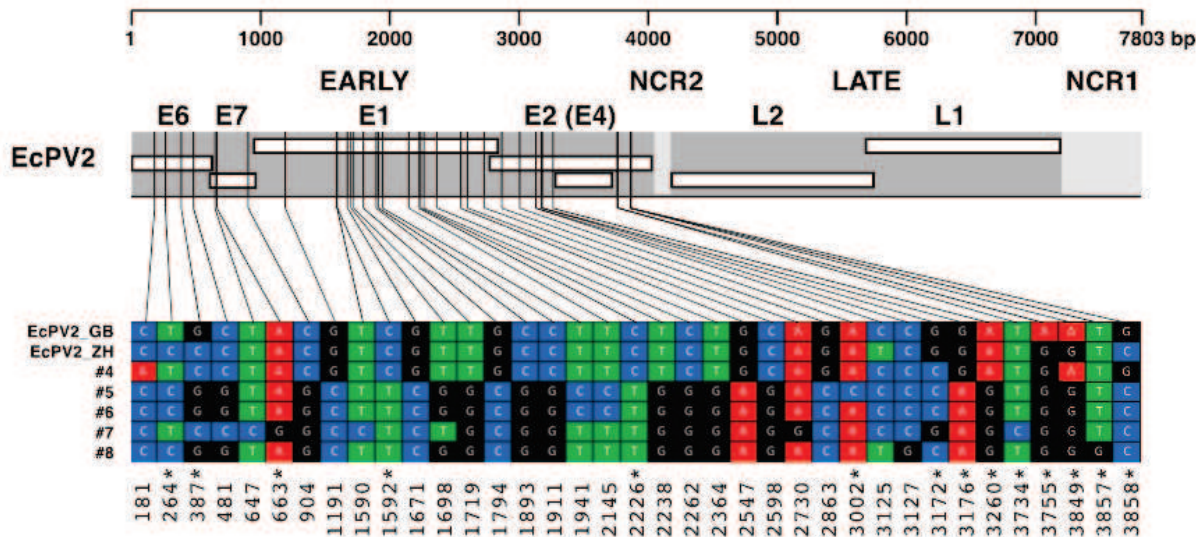


Figure 3: Sequence variations

The EcPV2 sequences detected in the five horses compared to the two published sequences are listed below the schematic view of the EcPV2 genome. Cytosin is shown in blue, Guanin in black, Thymin in green and Adenin in red. There were 37 sequence variations, thereof 23 were silent, while 14 were non synonymous point mutations. The positions of the sequence variations are written below and the nonsynonymous are marked with a star. Two of the non-synonymous point mutations were detected in the E6 at position 264 (Ile (ATC) --> Thr (ACC)) and 387 (Arg (AGG) --> Thr (ACG)), one in the E7 at position 663 (Thr (ACC) --> Ala (GCC)), two in the E1 at position 1592 (Ala (GCG) --> Val (GTG) and 2262 (His (CAC) --> Gln (CAG) and nine in the E2 at position 3002 (Ile (ATC) --> Leu (CTC)), 3172 (Gln (CAG) --> His (CAC)), 3176 (Val (GTC) --> Ile (ATC)), 3260 (Thr (ACA) --> Ala (GCA)), 3734 (Ser (TCT) --> Pro (CCT) 3755 (Asn (ACC) --> Asp (GAC)), 3849 (Lys (AAA) --> Arg (AGA)), 3858 (Cys (TGT) --> Ser (TCT)) and 3857 and 3858 both together (Cys (TGT) --> Ala (GCT))

.

3.3.2. Horse transcriptome

Our next aim was to evaluate differences in the host genes expression in EcPV2 infected SCC samples compared to healthy tissue samples. Differently expressed (DE) genes were characterized and used for further analysis to detect affected pathways, GO terms for biological processes and potential marker genes.

3.3.2.1. Gene counts positive over negative

The goal was to detect DE host genes in EcPV2 positive SCCs compared to negative healthy tissue samples. Therefore, the RNA-Seq reads were mapped to the horse genome and a two group comparison analysis of positive over negative samples was done. The mapping resulted in 13717 transcripts with counts above the linear threshold of 10. These counted reads were used for the analysis of DE genes. The gene transcript counts sorted by significance and fc (fc = $2^{|log_2 ratio|}$) are shown in Table 3. In the left part of the table, the number of DE genes according to the indicated p-value threshold and the corresponding false discovery rate (FDR) are listed. In the right part of the table, the numbers of DE genes depending to decreasing p-value thresholds and increasing fc thresholds are shown. As expected the number of DE genes decreased with the decreasing p-values, but there were still numerous genes with low p-value and high fc.

Table 3: Sorted significance and fold change of positive over negative gene counts.

	#significants	FDR
p < 0.1	4755	0.2882
p < 0.05	3610	0.19
p < 0.01	1957	0.06994
p < 0.001	937	0.01458
p < 1e-04	470	0.002849
p < 1e-05	257	0.0005194

	fc > 1	fc > 1.5	fc > 2	fc > 3	fc > 4	fc > 8	fc > 10
p < 0.1	4755	4755	3824	2244	1526	635	503
p < 0.05	3610	3610	3320	2065	1449	628	500
p < 0.01	1957	1957	1957	1566	1162	580	466
p < 0.001	937	937	937	907	771	432	353
p < 1e-04	470	470	470	469	448	296	247
p < 1e-05	257	257	257	257	254	190	165

The 1957 genes with p-values below 0.01 were used for most of the further analysis. All this genes had fold changes above 2 (≙log2 ratios >I1I) (Table 3). Their distribution depending on log2 ratios is shown in Figure 4. Numerous significant DE expressed genes (p<0.01) were quite high up or down regulated.

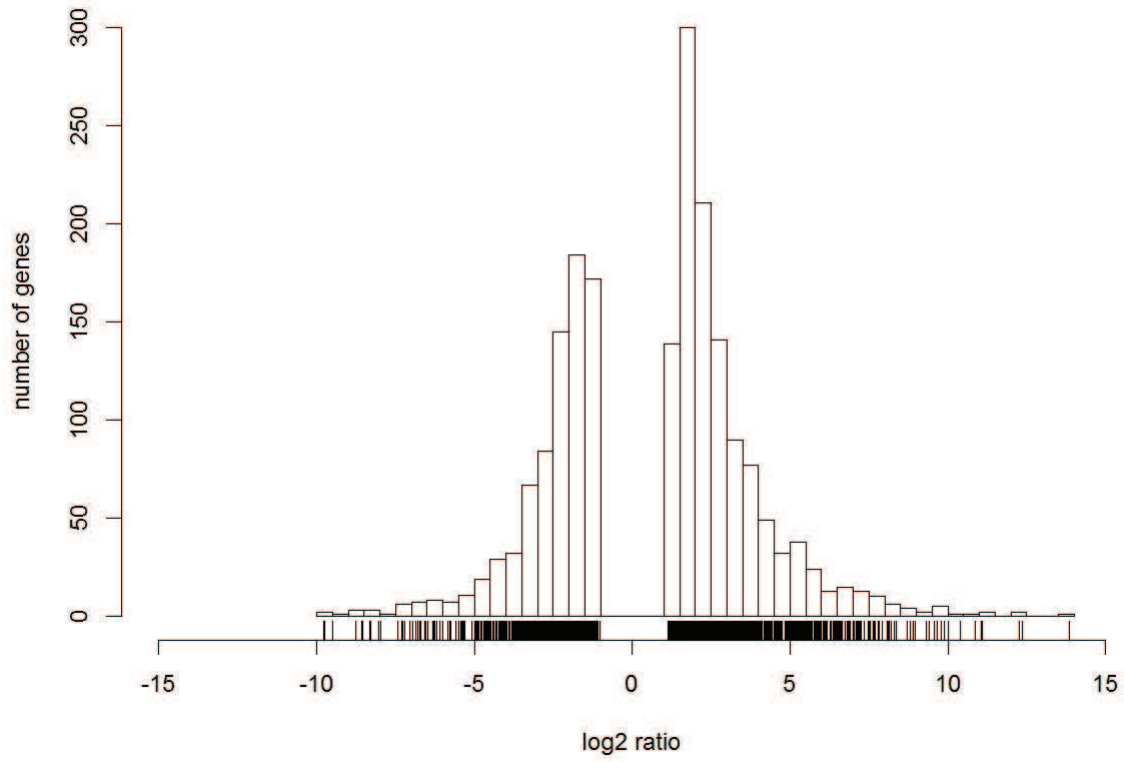


Figure 4: Histogram of log2 ratios p < 0.01

The histogram shows the distribution of the numbers of genes depending on their log2 ratios for the 1957 DE genes with p-values below 0.01. Log2 ratios below zero show the downregulated genes while above zero the upregulated genes are shown.

3.3.2.2. Clustering of significant genes

To compare the gene expression profile within the different samples a cluster analysis was done. The 1957 differently expressed genes with p-values below 0.01 and fc above 2 were used in an unsupervised cluster analysis. The cluster plot is shown in Figure 5. The clustering of the up and down regulated genes revealed specific patterns for negative and positive samples, respectively. However the expression pattern of sample #4 (Positive 1) differed from the patterns of the other positive samples.

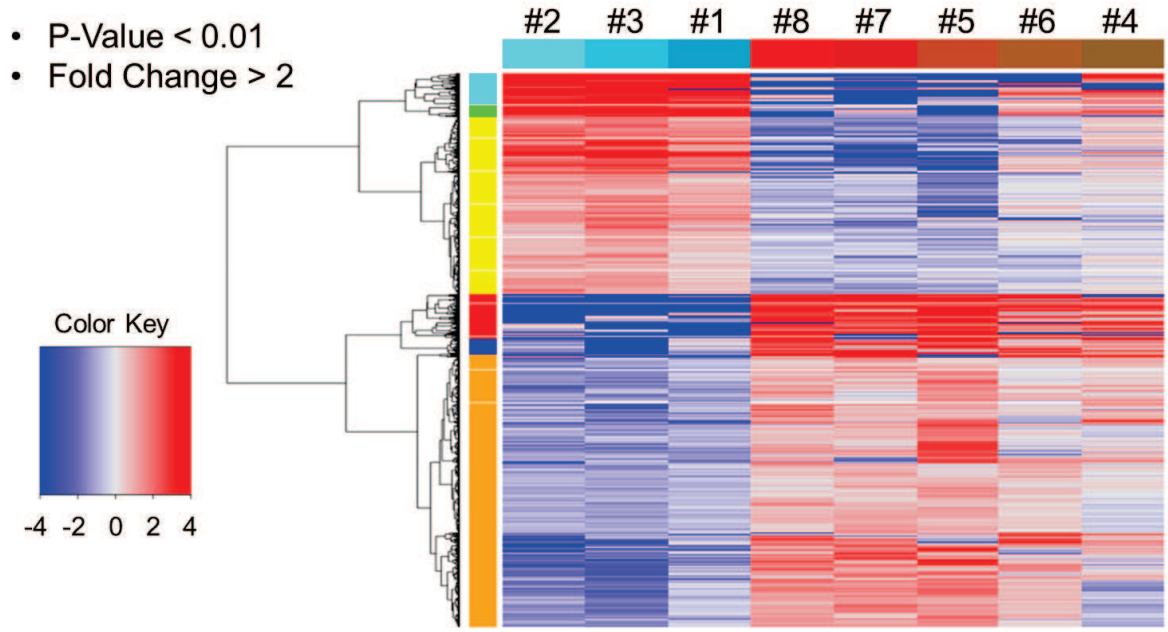


Figure 5: Clustered heatmap of 1957 significant genes

1957 genes with a significance threshold of p<0.01 and fc >2 were used for clustering of significant features. Upregulated genes are depicted in red, while downregulated in blue. The three EcPV2 negative samples (#1-3) are on the left side and the five positive samples (#4-8) on the right side.

3.3.2.3. KEGG Pathway analysis

For the detection of affected pathways, which might be involved in the development of EcPV2 associated SCC, DE host genes were used for KEGG pathway analysis by "GSEA", "STRING" AND "DAVID". A list of genes with a p-value <0.01 (1957) were used for pathway analysis by "DAVID" and "STRING", while for "GSEA" analysis all genes (13717) with corresponding fc-values were included. The significance of the most affected pathways detected by all the three programs is shown in Figure 6. We identified the cell cycle including DNA replication as significantly upregulated. Also extracellular processes including ECM receptor interaction and focal adhesion were recognized to be affected. This suggests that the disease might be driven by intracellular factors, which upregulate the mitosis rate and extracellular factors which affect interactions between the cells.

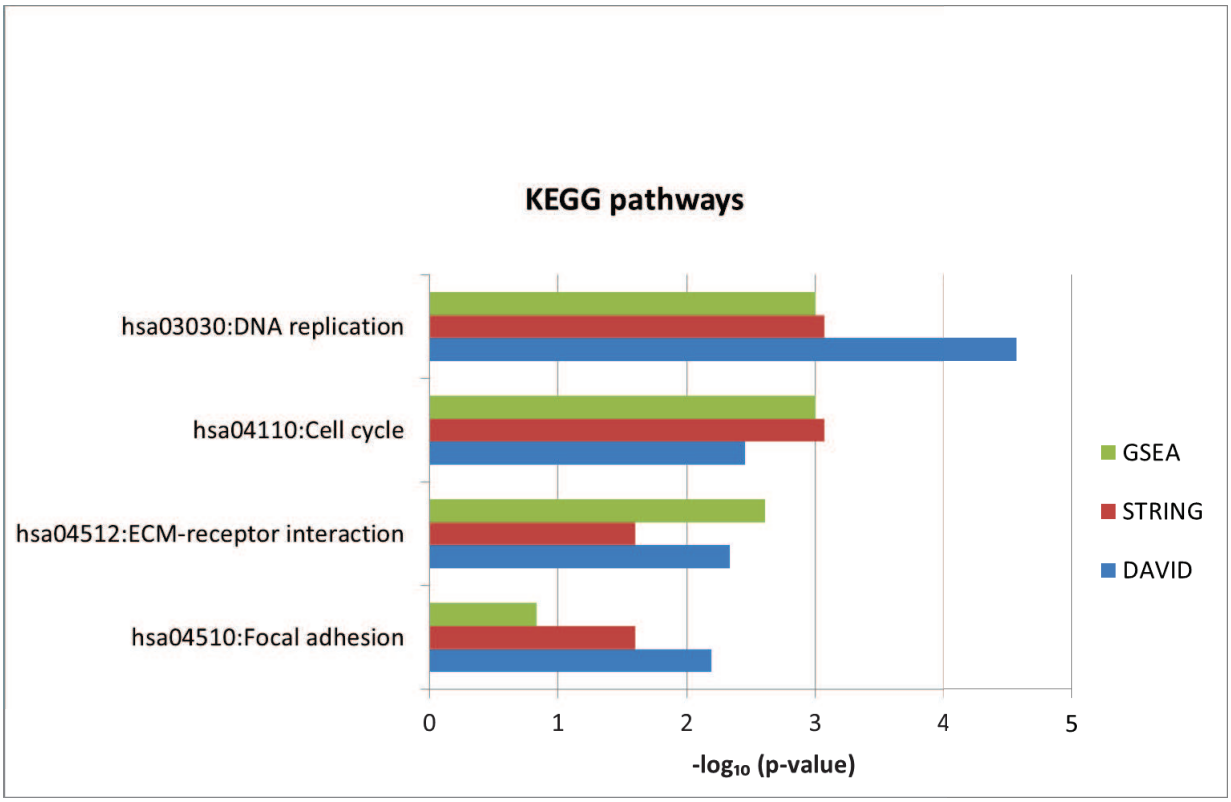
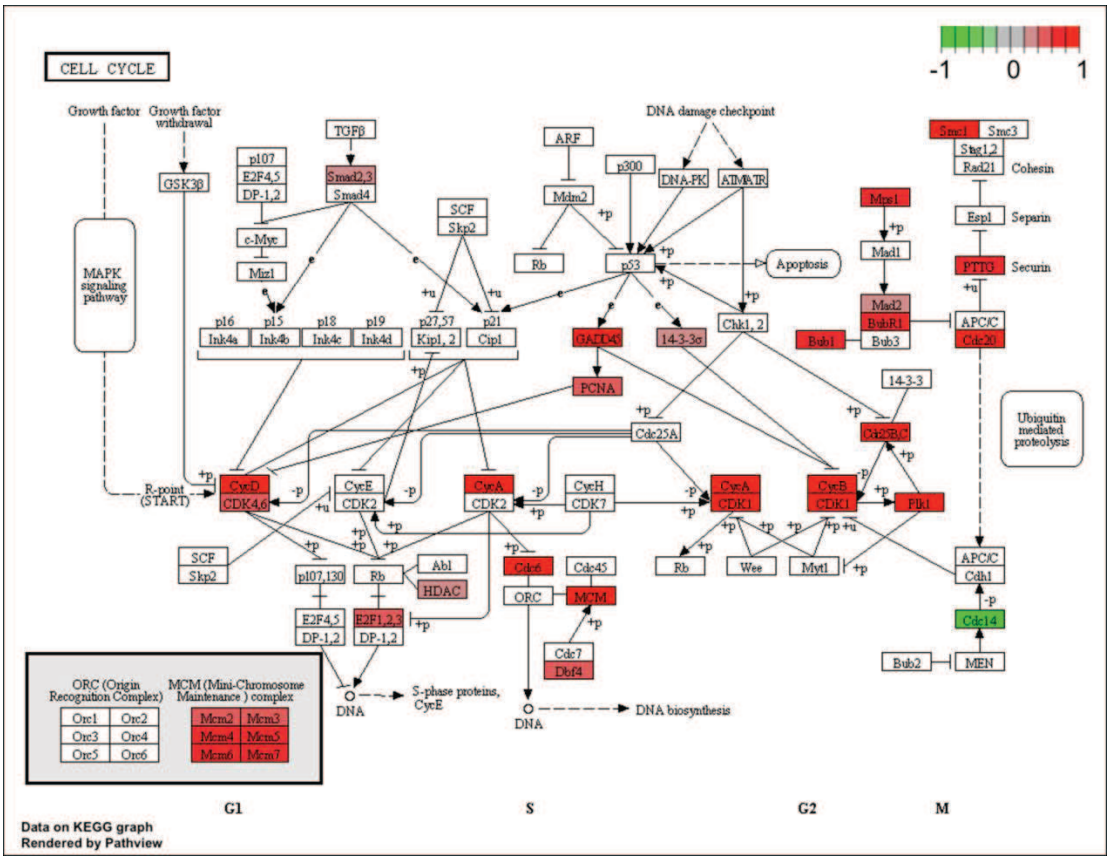


Figure 6: Significantly affected KEGG pathways by "GSEA", "STRING" AND "DAVID"

KEGG pathway analysis was done by "GSEA" (http://www.broadinstitute.org/gsea), "STRING" (http://string-db.org) and "DAVID" (http://david.abcc.ncifcrf.gov). For "GSEA" analysis all 13717 DE genes were included, while a list of 1957 DE genes with p-value <0.01 were used for pathway analysis by "DAVID" and "STRING", according to the different methods the programs are using. The four significantly affected pathways detected by all the three programs are plotted against the $-log_{10}$ (p-values) obtained from the different programs. (p-values of "GSEA" analysis below 0,001 are shown as 3)

To further characterize the affected genes within the pathways the cell cycle and the focal adhesion pathway, the ratios were projected onto the KEGG pathway maps and shown in Figure 7. Up or down regulated genes and the intensity of regulation are distinguished by the colours given to the genes. In the cell cycle pathway the majority of genes were upregulated while there were up as well as down regulated genes in the foal adhesion pathway.



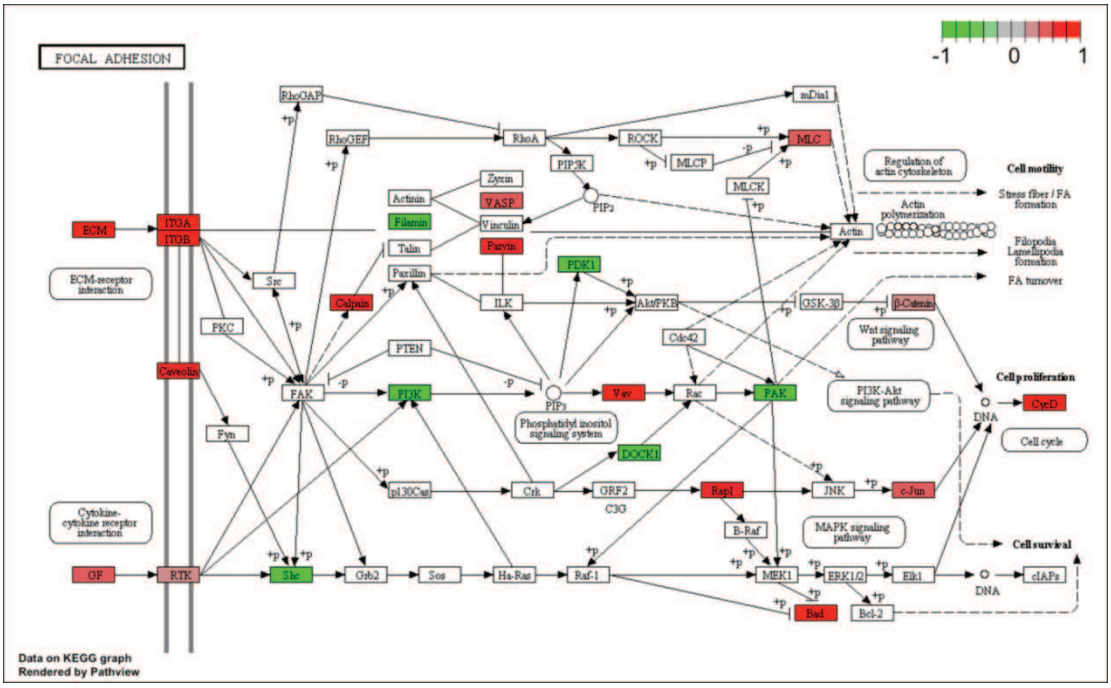


Figure 7: Affected genes KEGG Pathway Cell Cycle and Focal adhesion

Schematic graph of KEGG Pathway "cell cycle" and "focal adhesion". Up regulated genes are shown in red and down regulated in green.

3.3.2.4. GO term BP analysis

For the detection of affected biological processes a GO term BP analysis was also done by "GSEA", "STRING" AND "DAVID" with the identical lists of genes as used for the KEGG pathway analysis. Figure 8 shows the significance of the seven most enriched GO terms detected by all the three programs. This result confirmed the observation of the pathway analyses, the cell cycle seems to be significantly regulated, especially the Mitosis.

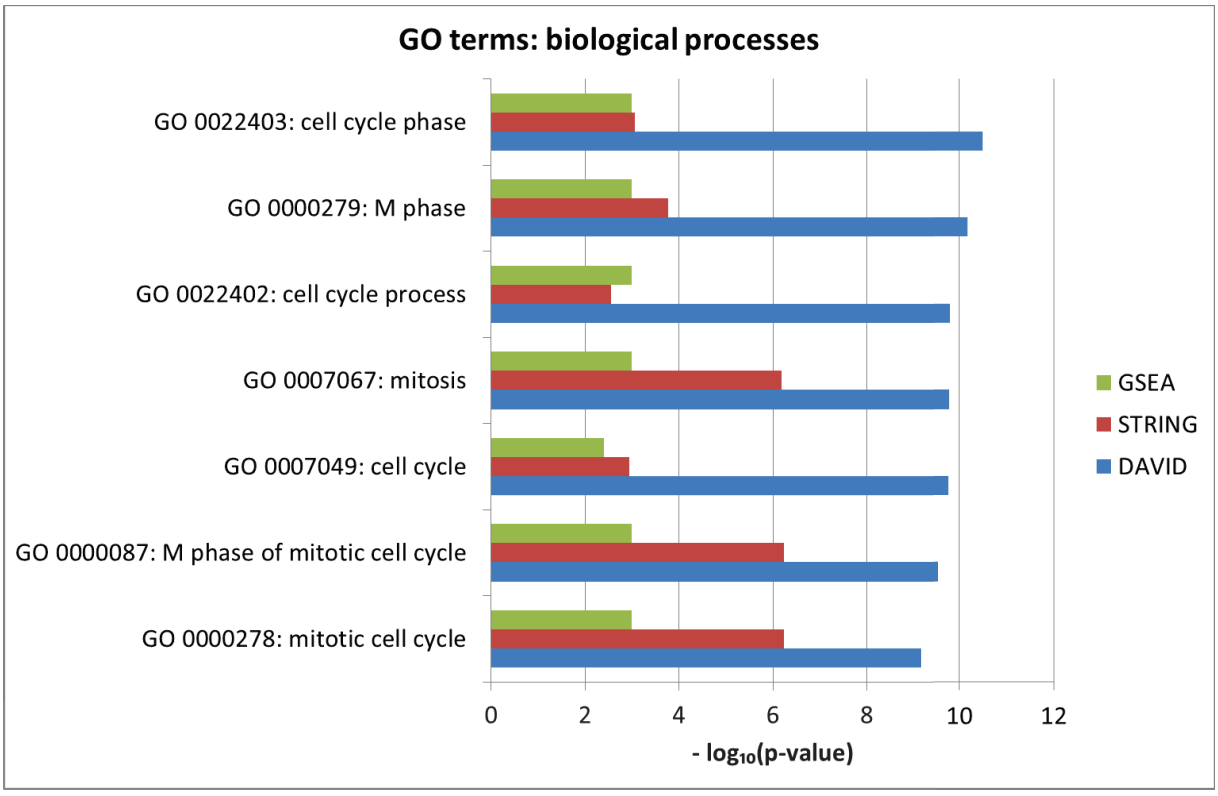


Figure 8: GO terms biological processes analysis

The $-log_{10}$ (p-values) of the seven most significantly enriched GO terms calculated by three different programs are plotted. (A "GSEA" $-log_{10}$ (p-values) at 3 stands for p-value < 0.001)

To visualize the affected biological processes, an enrichment map of the GO term BP identified by "GSEA" analysis of differentially expressed host genes was constructed using "CYTOSCAPE". As seen in Figure 8 the cell cycle including DNA metabolic process is intensely affected. This confirmed the results of the KEGG pathway analysis. In addition, RNA processing and response to stress were detected as significantly regulated as well.

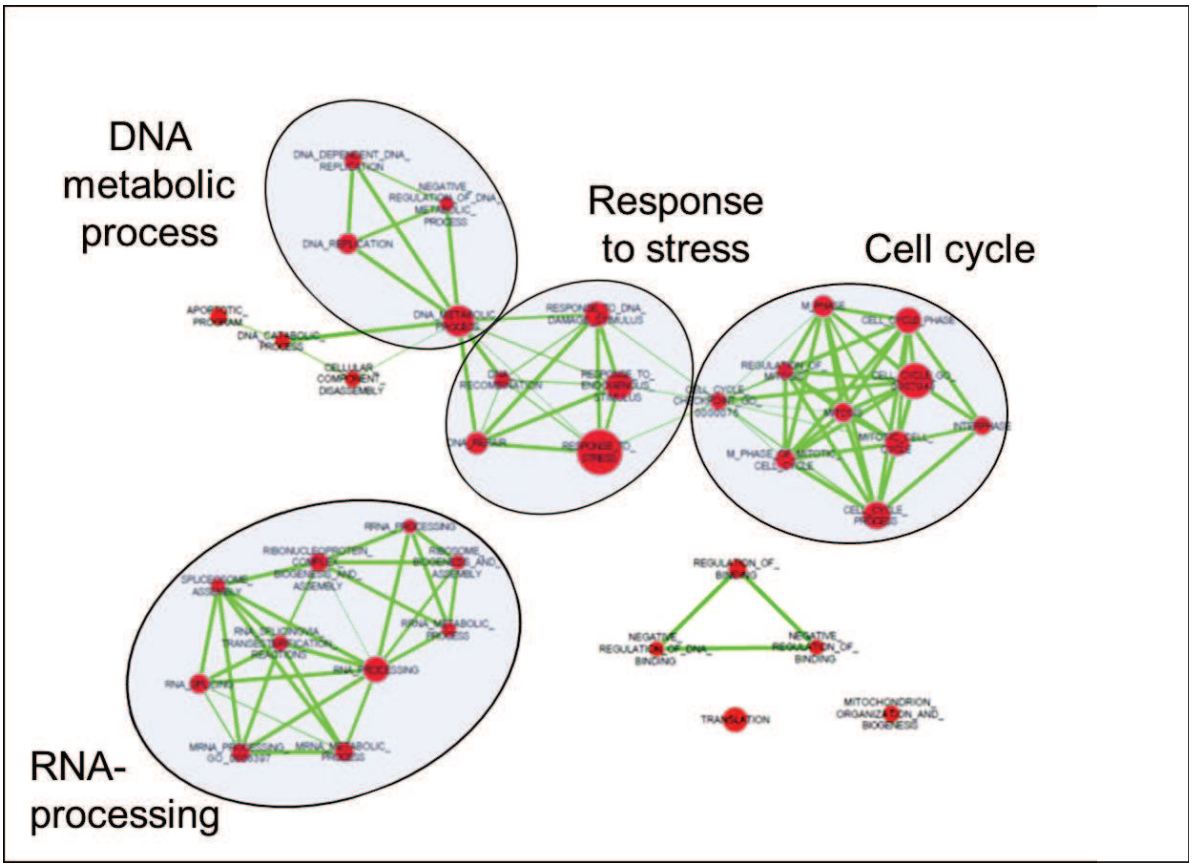


Figure 9: Enrichment Map of the "GSEA" GO terms biological processes analysis

An Enrichment Map of the "GSEA" GO terms BP analysis was constructed using "CYTOSCAPE" (http://www.cytoscape.org/). Nodes represent individual GO terms and their sizes correspond to the number of assigned genes. Edges represent common genes in the connected nodes and the thickness the numbers of common genes. Cell cycle, RNA processing, response to stress and DNA metabolic process are the most affected biological processes.

3.3.2.5. Potential marker genes

Molecular diagnostics for SCC based on elevated gene expression demands the identification of marker genes. Therefore, one aim of this study was to define such potential marker genes. The hundred most up regulated genes of our study were compared by "EXPRESSION BLAST" to gene expression studies published on GEO. The analysis was performed twice, either with "carcinoma" as keyword preselected datasets or with non-preselected datasets. The results are shown in Figure 10.

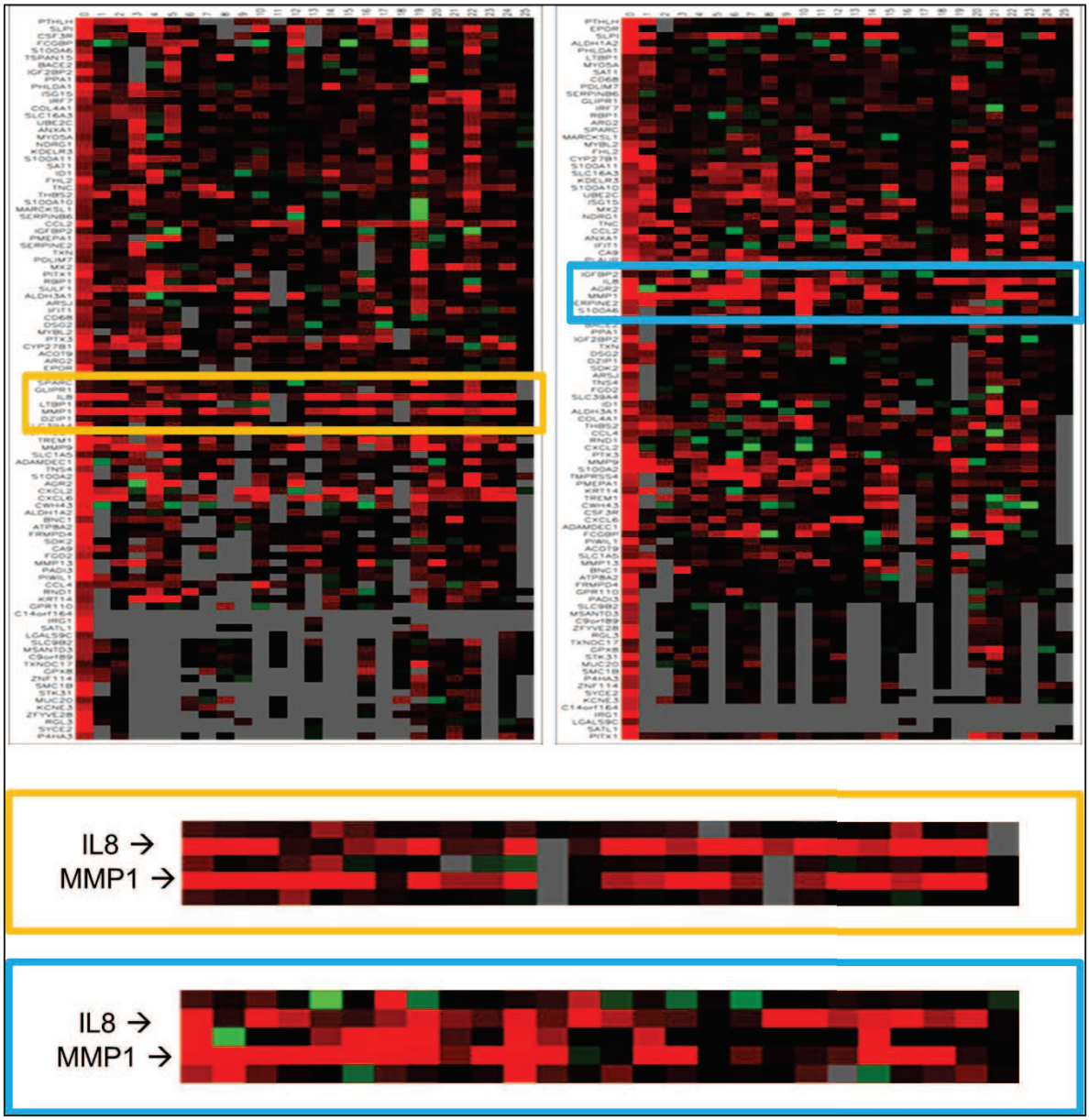


Figure 10: Heatmap Top 100 upregulated genes compared to other studies

*To detect potential marker genes a gene list of the TOP 100 up regulated genes with the lowest p-values were compared via "*EXPRESSIONBLAST*" (status 07/2014) to other gene expression studies. For the heatmap on the left side studies were chosen without pre selection and on the right side "carcinoma" was used as a keyword to look especially for carcinoma studies. On the left hand side of the plots, the 100 genes are listed and on the top the 25 most similar studies depending on these genes are shown (The corresponding studies are shown in Supplementary Table 3 and 4). Number 0 is the genelist from this study. Upregulated genes are shown in red and downregulated genes in green. In both cases IL8 and MMP1 seem to be intense up regulated in many other studies.*

The analyses with both settings revealed Interleukin 8 (*IL8)* and Matrix Metalloproteinase 1 (*MMP1)* as up regulated genes in ours as well in many other studies. The list of the 100 most upregulated genes sorted by p-values is shown in Supplementary Table 2. Both genes revealed high log2 ratios (IL8: 13.85 and MMP1: 9.43) which resulted in a five or three digit fold fc, respectively (IL8: 14766 and MMP1: 690). MMP1 is an interstitial collagenase, which is involved in the breakdown of extracellular matrix. IL8 is a chemokine, which induces chemotaxis and is also a factor for angiogenesis. Previous studies have reported these genes to be upregulated in PV associated cancers (Walker et al. 2011; Hsiao et al. 2013; Rajkumar et al. 2011; Mosseri et al. 2014; Yuan et al. 2010; Yuan et al. 2008; Regezi et al. 2002; Akgül et al. 2005; Ryzhakova, Solov'eva 2013) and therefore *MMP1* and *IL8* might be useful as biomarkers for the development of SCC in EcPV-2 infected tissue.

4. Discussion

The knowledge about EcPV2 and its possible association to develop SCC is still limited. Therefore, EcPV2 gene expression and the changes in host gene expression of affected tissue were characterized on a transcriptional level for the first time by RNA-Seq. By analysis of the viral transcriptome we could quantify the transcription levels and the splicing events of viral mRNAs. By analysis of the host's transcriptome, we could confirm our hypothesis that the gene expression profile of EcPV2 infected tissue is altered as compared to healthy skin. Our evaluation of these differences resulted in 1957 DE host genes. These DE host genes were further characterized and revealed biological processes as cell cycle, focal adhesion and RNA processing to be the most affected ones. Additionally IL8 and MMP1 were defined as potential marker genes that could be used for assessing future development of SCCs on the affected penis of horses.

4.1. Viral transcriptome

4.1.1. Viral reads

In the first part of this study, the whole transcriptome of EcPV2 was assessed. The analysis of the reads mapped to the viral genome revealed comparable gene expression and splicing in all EcPV2-PCR positive samples, while no reads were detected in the EcPV2-PCR negative samples. For all positive samples, the gene expression profile of EcPV-2 showed high expression of the oncogenes E6 and E7 and other early genes in particular E4 and E2. While low expression of E1 and almost no expression of the late genes L1 and L2 were detected (Figure 1). Similar viral gene expression profiles are reported in high risk HPV cancers. The oncogenes E6 and E7 get deregulated during an HPV16 infection, sometimes due to the integration of the viral DNA into the host genome, which might lead to the cancer progression. The viral life cycle cannot be completed in such a situation and therefore there is no expression of the late capsid proteins (Doorbar 2005; Zur Hausen 2002). RNA-Seq analysis of HPV16 positive cervical SCCs and HPV58 positive cervical intraepithelial neoplasias (CIN) 2/3 showed similar expression results, except for the E5 gene which is not present in EcPV2 (Chen et al. 2014; Yang et al. 2013). The comparison to human skin cancer failed as there were no viral reads detected in RNA-Seq analysis of papillomavirus positive skin cancers, suggesting that the virus is not active in

human cutaneous SCCs (Arron et al. 2011). No complete expression study covering all EcPV2 genes in positive SCCs is published yet. Indeed, Van den Top et al. determined the expression of the viral genes E2, E6 and L1 in non-invasive and invasive lesions. However, no differences in viral mRNA expression between the different lesions could be detected. They explained their findings by a possible “hit and run” action of EcPV2, suggesting that EcPV2 might lead to a penile intraepithelial neoplasia or a papilloma in the horse, while the followed progression to SCC is triggered by other mechanisms (Van den Top, J. G. B. et al. 2014). Therefore a longitudinal study of the gene expression in the equine penis papilloma might be very helpful. However, since the detailed expression levels were not presented in the publication, no further conclusion can be drawn. Likewise, Sykora et al. checked for E6 and E1 mRNA expression in 8 SCC samples. They could detect E6 mRNA in all the samples, while E1 was only detected in three of the samples. They interpreted their results by probable splicing events and reduced primer binding affinity in the PCR assay (Sykora et al. 2012). The low E1 expression compared to the higher E6 expression is in agreement with our RNA-Seq results. In conclusion, we describe here the first viral gene expression results covering the entire EcPV2 genome. Due to the small sample size of our study, the results should be confirmed in a broader range of samples.

4.1.2. Splice junctions of EcPV2

Gene expression regulation of PV takes place on the transcriptional level as well as on the level of pre-mRNA processing. Much is known about the role of alternative splicing for the regulation of HPV16 RNA processing during the early and late stages of infection, while there exists not much information about EcPV2 pre-mRNA processing yet (Johansson, Schwartz 2013). Transcript maps for BPV1 and several HPVs are available on PAVE (http://pave.niaid.nih.gov/) (van Doorslaer et al. 2013), but not for EcPVs. Therefore we screened the RNA-Seq data for detection of splice events in EcPV2 infected tissue. 21 EcPV2 splice junctions on intron-spanning reads could be detected. Thereof the six most frequent splice junctions (with more than 50 reads in total over all samples) are shown in Figure 2. The most common junctions are between E1^E2/E4, E1^E2 and E2/E4^E2/E4, while most prevalent are donors at nucleotide position 964 and 1368 and acceptors at 3370 and 2721. Two of the products are predicted spliced genes (in detail E1^E4 (943..964, 3370..3767) and

E8^E2 (1331..1368, 3370..4018)) of the EcPV2 genome by PAVE (http://pave.niaid.nih.gov/) (van Doorslaer et al. 2013). All acceptors and donors, except of the donor at nucleotide position 1368, were splice sites *in silico* predicted by NetGene2 (http://www.cbs.dtu.dk/services/NetGene2/) (Brunak et al. 1991; Hebsgaard et al. 1996). The positions of the most prevalent donors at nucleotide position 964 and 1368 and acceptors at 3370 and 2721 relative to the ORFs are similar to published splice sites of diverse HPVs and BPV1 genomes (http://pave.niaid.nih.gov/) (Johansson, Schwartz 2013; van Doorslaer et al. 2013) and the coding potentials of the mature mRNAs might be comparable.

The most frequently detected splice junction (E1^E4 964-3370) has also been described for many HPVs and BPV1, thus, representing splice donor and acceptor sites that are conserved across different PV types (http://pave.niaid.nih.gov/) (van Doorslaer et al. 2013). It is known that E4 gene products are expressed from these E1^E4 spliced mRNA (Doorbar 2013) and therefore this splice junction is in the case of EcPV2 also the most likely one used to express the E4 gene. The E4 protein has been suggested to serve as a biomarker of active virus infection, and, in the case of high-risk HPV types, also disease severity. It may account for as much as 30% of total lesional protein content. Therefore, it can easily be visualised in biopsy material by immunostaining. Functionally, the E4 protein may facilitate efficient virus release and transmission, while, during early phases, also contributing to genome amplification-efficiency (Doorbar 2013). In the future, detection of EcPV2 E4 protein in equine lesions should be established to further assess the virus's contribution equine carcinomas.

The second and fourth most frequent splice junctions (E1^E2 964-2720 and E1^E2 1368-2720) have the same splice acceptor (2720) in common which is also a conserved splice acceptor in other PVs (http://pave.niaid.nih.gov/) (van Doorslaer et al. 2013). It is known that splicing to this acceptor generates mRNAs for expression of full length E2 protein (Johansson, Schwartz 2013), suggesting that this might be the case for EcPV2 as well. Comparable splice junctions like the first (964-2721) are present in almost all HPVs, while the second one (1368-2721) is only detected in BPV1 and a few HPVs (HPV5, HPV11, HPV47) (http://pave.niaid.nih.gov/) (van Doorslaer et al. 2013). The E2 proteins function as the main transcriptional regulators of PVs, by recruiting cellular factors to the viral genomes, which activate or repress

transcriptional processes. Besides the full length E2 transcript, all PVS have the potential to encode shorter E2 forms by spliced messages that link sequences from an alternative reading frame in the E1 region of the genome (designated E8) to the C-terminus of E2 (McBride 2013). This splice junction (1368-3370) named E8^E2 is also present in the EcPV2 transcripts. This shorter form of E2 functions as a repressor of viral transcription and replication (McBride 2013).

Not much information is available for the two short splice junctions E2/E4^E2/E4 (3197-3370 and 3528-3653). A similar junction like our first E2/E4^E2/E4 (3197-3369) was detected in HPV18. There, the splice donor site is located in the 5′ part of E2, while the splice acceptor site is the one used for E1^E4, like in our detected EcPV2 junction. The level of E2^E4 spliced transcripts seems to increase during keratinocyte differentiation. These E2^E4 fusion proteins are cytoplasmic and can probably induce cell death. Though the functions of these proteins are not completely understood, it is believed, that they could functionally mimic the E2/E1^E4 interaction and could also exhibit new functions that might be important for the viral life cycle and pathophysiology (Tan et al. 2012). A similar junction like the other short E2/E4^E2/E4 splice junction (3528-3653) was only detected once in HPV5. There, for the first time an internal splice site in the E4 ORF was recognized in putative L2 messages. By PCR with E1/L2 and NCR/L2 specific primer pairs, a peculiar E4 internal splice site was detected. While the transcript starting from the NCR seem to encode for a putative L2 message, the transcript starting from E1 might encode for L2 and truncated E1^E4 protein with unknown function (Haller et al. 1995).

It has to be kept in mind, that by RNA-Seq only small reads of mRNA are detected. Consequently, we have just information about the splice junctions and no information about the whole generated mRNA transcript. Furthermore, we do not know the combination of more than one splice event during processing of pre-mRNA to a certain mRNA. Therefore by comparing to the literature, just suggestions about the coding potential of these mRNAs could be done. To figure out, which whole mRNA transcripts are generated by EcPV2 and what the functions of these transcripts and their translation products are, further experiments are needed. However, in this study for the first time splice junctions of EcPV2 were detected. The four most frequent are in common with other PVs and have predicted functions, which might be conferrable to EcPV2. Two further small splice junctions, which seem to be not so much

conserved in other PVS, were recognized in the E2 and E4 region. To understand the functions of this alternative splicing further studies are warranted. Finally, one may consider that also introns may be of biological relevance, for example in the form of micro RNA as fine-tuning regulators of gene expression. However, presently no bioinformatical evidence in this context could be generated.

4.2. Horse transcriptome

4.2.1. Biological processes

In the second part of this study, the transcriptomes of EcPV2 affected and non-affected horse tissues were compared and identified DE genes were analysed to reveal the involved pathomechanisms driving this disease. Enrichment analysis for biological processes was performed by several programs. The most affected processes are on the one hand the cell cycle and RNA processing and on the other hand ECM receptor interaction and focal adhesion.

The most significantly influenced biological process is the cell cycle, in particular the mitosis (Figure 6, 7, 8, 9). As a higher mitosis rate is expected in cancerous tissue, these results are explainable by the cancerogenesis, as well as in some aspects by the virus or both. It is well known, that the cell cycle is upregulated during PV cancer development. The cell cycle genes have been described to be one of the main drivers of cervical cancer (Mine et al. 2013). Furthermore, by expression analysis using microarrays, genes involved in cell cycle and cell proliferation were detected to be dysregulated in BPV1 associated equine sarcoid (Yuan et al. 2008). Of note, it seems that cell cycle and DNA damage response are much stronger affected by high-risk HPV as compared to low-risk HPV (Santegoets, Lindy A M et al. 2012). As EcPV2 seems to be an “equine high-risk PV” this might be the reason for upregulated cell cycle similar to HPV infections. In particular the interaction of the HPV oncogenes E6 and E7 with regulators of the cell cycle are known to be the cause of cervical cancer. (Clarke, Chetty 2001). As our study shows a high expression of the EcPV2 oncogenes E6 and E7 (Figure 1), it might also explain the upregulation of the cell cycle. The oncoproteins E6 and E7 induce immortalization of cells partly through their inhibitory effects on tumour suppressor proteins pRb and p53. The pRb’s growth-inhibitory role allows release of E2F transcription factors, from which some associated genes are also upregulated in our study, especially *E2F1* (Figure 7;

Supplementary Table 1). Increased E2F grants limitless proliferative potential by allowing expression of products such as cyclins A, E, and B, dihydrofolate reductase, and DNA polymerase which provokes various stages of the cell cycle and leads to override the G1–S and G2–M cell cycle checkpoints. Upregulation of a gene encoding for cyclin A, (in detail *CCNA2*) was also detected in this study (Figure 7 and Supplementary Table 1) and it is known, that cyclin A promotes anchorage-independent growth, which facilitates tissue invasion and tumour spread (Clarke, Chetty 2001). In general the cyclin-dependent protein kinases (CDKs) are known to be the key regulators of the eukaryotic cell cycle. The successive activation of various CDKs is required for the passage through the cell cycle. They build complexes with cyclin, which promote progression towards DNA replication. The CDK/cyclin complexes phosphorylate proteins required for the activation of genes involved in DNA synthesis, DNA replication and the entry of cells into mitosis (Nigg 1995). CDK2, CDK3, CDK4 and CDK6 drive cells through interphase and CDK1 is needed to proceed the cell through mitosis (Malumbres, Barbacid 2005). Of note, CDK1 seems to be the only essential cell cycle CDK. After phosphorylation of the retinoblastoma protein pRb and the expression of genes that are regulated by E2F transcription factors, CDK1 can bind to all cyclins and hence can execute all the events that are required to drive cell division in the absence of interphase CDKs (Santamaría et al. 2007). Genes encoding for CDKs and cyclins, in particular *CDK1* and *CDK4*, as well as for Cyclin A, B and D (in particular *CCNA2, CCNB1, CCNB2, CCNB3, CCND2*), are shown to be up regulated in EcPV2 infected cancer tissue samples (Figure 7; Supplementary Table 1). This might be a possible explanation for the upregulated cell cycle. However, many other factors like phosphorylation play a role in these biological processes and posttranscriptional changes, as phosphorylation, were not addressed by this study. The exact mechanisms involved in the detected processes have to be studied in more detail in future experiments.

Several post-transcriptional regulations, in particular RNA processing, were detected as significantly dysregulated in the EcPV2 positive horses (Figure 9). This might be explainable through the presence of the PV. Since PV genomes are small and compact, post-transcriptional regulation is important to complete their life cycle and, regulation of RNA processing in the nucleus is essential for the regulation of viral gene expression (Graham 2010). Such changes are also observed in equine BPV1

associated sarcoids, where a number of genes involved in transcription regulation and RNA processing and metabolism are deregulated (Yuan et al. 2008).

Besides these intracellular factors described above, extracellular factors were detected by the pathway analysis, too. Two significantly affected pathways, ECM receptor interaction and focal adhesion, were identified (Figure 6). Deregulation of these pathways is beneficial to support the survival, growth, and invasion of cancer cells and is an essential player at various stages of the carcinogenic process (Lu et al. 2012). Besides the regulation by the CDKs, the balance between cell cycle arrest and cell proliferation is controlled by the extracellular matrix and contact inhibition (Gérard, Goldbeter 2014). Genes involved in cell adhesion, motility, and integrity are also affected in BPV1 positive equine sarcoids (Yuan et al. 2008). The influence of PV infection on the focal adhesion was already shown by several studies. The E6 protein was reported to induce a strong modulation of focal adhesion through the degradation of TAp63β protein (Ben Khalifa et al. 2011), by the interaction with paxicillin and FAK (Tong, Howley 1997; Sarode, Sarode 2014) or by modulation of β1-integrin signalling (Holloway, Storey 2014). From these genes only these for the integrins, which were described as elevated in non-melanoma skin cancer, were affected in our study (Figure 7). The most affected ones are *ITGA3, ITGA6, ITGAX, ITGB1, ITGB2 and ITGB4* (Supplementary Table 1). Of note, the specific interactions between cells and the ECM are mainly mediated by integrins, which function as mechanoreceptors and provide a force-transmitting physical link between the ECM and the cytoskeleton (http://www.genome.jp/kegg-bin/show_pathway?hsa04512) (Kanehisa, Goto 2000; Kanehisa et al. 2014). The integrins seem to have a key role in the ECM receptor interaction and focal adhesion pathway. As shown in Figure 7, the increased ECM receptor interaction has a direct influence on the focal adhesion pathway, as well as the cytokine receptor interaction. Finally, the disturbances of the focal adhesion pathway seen in our study result in an up regulation of the cell cycle by increased proliferation (in particular via Cyclin D). In conclusion all the detected extra- and intracellular factors in the pathway and biological processes analysis are interconnected. These changes might be induced by the EcPV2 infection and might also be caused by or resulting in the cancer development.

4.2.2. Potential marker genes

Another aim of this study was the definition of potential marker genes for the development of SCC in PV affected horses, which could be useful for diagnosis and prognosis. Two potential genes were detected: *MMP1* and *IL8* (Figure 10). Both genes were recognized to be increased in many other cancer studies (Supplementary Table 3 and 4) and their proteins are known to have an influence on the ECM receptor interaction and thus focal adhesion.

MMP1 is part of the MMP (matrix metalloproteinase) family, which is involved in the breakdown of extracellular matrix in normal physiological processes, as well as in disease processes. The function of MMP1 is degradation of collagen type I, II and III (http://www.ncbi.nlm.nih.gov/gene/4312) (Pruitt et al. 2014). It has been demonstrated, that the MMPs play a causal role in tumour cell invasion by initiating degradation of basement membrane and extracellular matrix (Westermarck, Kahari 1999). Some studies show an up regulation of the *MMP1* gene in PV associated cancers. *MMP1* gene expression is essentially increased by HPV E7 oncoprotein in most HPV 16 and 18 positive cell lines and was also described to be upregulated in cervical cancers (Ryzhakova, Solov'eva 2013; Rajkumar et al. 2011). In HPV8 the E7 protein has been shown to stimulate the overexpression of MMP1 protein, too. This overexpression promotes an invasive phenotype by degradation of the basal membrane and ECM, which then allows invasion of human keratinocytes into the dermis (Akgül et al. 2005). Of note, MMP1 gene and protein expression is not or only moderately expressed in HPV positive dysplastic oral warts while it is highly expressed in oral HPV positive SCCs (Regezi et al. 2002). This might allow to define this gene as a prognosticator. Interestingly, it was shown that BPV1 also upregulates *MMP1* gene expression in horses, while the MMP1 protein is essential for the transformation of sarcoid fibroblasts (Yuan et al. 2008; Yuan et al. 2010) and it was also already suggested that there is a causal correlation between MMP1 protein expression and the local aggressiveness of sarcoids regardless of the clinical type (Mosseri et al. 2014). These findings indicate that MMP1 might have an important impact on the formation of SCCs and might be a potential biomarker for the development of EcPV2 associated penile SCCs.

IL8 also known as CXCL8 (chemokine (C-X-C motif) ligand 8) is a member of the CXC chemokine family. IL8 is one of the major mediators of the inflammatory

response and functions as a chemoattractant, but it is also a potent angiogenic factor (http://www.ncbi.nlm.nih.gov/gene/3576) (Pruitt et al. 2014). Elevated expression of IL8 has been characterized in cancer cells, endothelial cells, infiltrating neutrophils, and tumour-associated macrophages. Therefore IL8 might function as a significant regulatory factor within the tumour microenvironment. Furthermore, inhibiting the effects of IL-8 signaling was suggested as a therapeutic option for cancer (Waugh, David J J, Wilson 2008). Interestingly, the functions of IL8 and the MMPs seem to be connected since IL-8 directly enhances endothelial cell survival, proliferation, MMP production and regulated angiogenesis (Li et al. 2003). Particularly the up regulation of MMP-2 and MMP-9 expression and activity is induced by IL8 (Shiau et al. 2013). Of note, the *IL8* gene is known to be upregulated amongst other genes in cervical cancers (Rajkumar et al. 2011). The expression of E6 and E7 proteins seems to come along with raised IL8 expression during PV infections (Walker et al. 2011; Hsiao et al. 2013). As increased IL8 signaling is indeed reported in numerous solid tumours during the most advanced stages of disease, (Waugh, David J J, Wilson 2008) it might also be useful as a potential marker for development of SCC.

In conclusion we detected two potential candidates, which could serve as marker genes for prognosis and diagnosis. However as there are no virus negative SCC horses and no healthy EcPV2 horses included in our study, it cannot be distinguished if the changes derive from the viral infection or the cancer development or both. The identification of the potential marker genes as well as all the other results need to be validated in a broader range of samples and confirmed by different methods.

In this study we gained valuable new information about EcPV2 and its influence on the host's gene expression. In summary we could identify the mainly affected biological processes and potential marker genes. Furthermore, we could quantify the transcription from the viral genome and the splicing events of some viral mRNAs. These results will be the basis to study EcPV2 and its associated cell changes in more detail.

References

Akgül, Baki; García-Escudero, Ramón; Ghali, Lucy; Pfister, Herbert J.; Fuchs, Pawel G.; Navsaria, Harshad; Storey, Alan (2005): The E7 protein of cutaneous human papillomavirus type 8 causes invasion of human keratinocytes into the dermis in organotypic cultures of skin. In *Cancer Res.* 65 (6), pp. 2216–2223. DOI: 10.1158/0008-5472.CAN-04-1952.

Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. (1990): Basic local alignment search tool. In *J Mol Biol* 215 (3), pp. 403–410. DOI: 10.1016/S0022-2836(05)80360-2.

Arron, Sarah T.; Ruby, J. Graham; Dybbro, Eric; Ganem, Don; Derisi, Joseph L. (2011): Transcriptome sequencing demonstrates that human papillomavirus is not active in cutaneous squamous cell carcinoma. In *The Journal of investigative dermatology* 131 (8), pp. 1745–1753. DOI: 10.1038/jid.2011.91.

Ben Khalifa, Youcef; Teissier, Sébastien; Tan, Meng-Kwang Marcus; Phan, Quang Tien; Daynac, Mathieu; Wong, Wei Qi; Thierry, Françoise (2011): The human papillomavirus E6 oncogene represses a cell adhesion pathway and disrupts focal adhesion through degradation of TAp63β upon transformation. In *PLoS pathogens* 7 (9), pp. e1002256. DOI: 10.1371/journal.ppat.1002256.

Bogaert, Lies; Willemsen, Anouk; Vanderstraeten, Eva; Bracho, Maria A.; Baere, Cindy de; Bravo, Ignacio G.; Martens, Ann (2012): EcPV2 DNA in equine genital squamous cell carcinomas and normal genital mucosa. In *Vet. Microbiol.* 158 (1-2), pp. 33–41. DOI: 10.1016/j.vetmic.2012.02.005.

Brunak, S.; Engelbrecht, J.; Knudsen, S. (1991): Prediction of human mRNA donor and acceptor sites from the DNA sequence. In *J Mol Biol* 220 (1), pp. 49–65.

Chen, Jinmiao; Xue, Yuezhen; Poidinger, Michael; Lim, Thimothy; Chew, Sung Hock; Pang, Chai Ling et al. (2014): Mapping of HPV transcripts in four human cervical lesions using RNAseq suggests quantitative rearrangements during carcinogenic progression. In *Virology* 462–463, pp. 14–24. DOI: 10.1016/j.virol.2014.05.026.

Clarke, B.; Chetty, R. (2001): Cell cycle aberrations in the pathogenesis of squamous cell carcinoma of the uterine cervix. In *Gynecologic Oncology* 82 (2), pp. 238–246. DOI: 10.1006/gyno.2001.6306.

Cline, Melissa S.; Smoot, Michael; Cerami, Ethan; Kuchinsky, Allan; Landys, Nerius; Workman, Chris et al. (2007): Integration of biological networks and gene expression data using Cytoscape. In *Nature protocols* 2 (10), pp. 2366–2382. DOI: 10.1038/nprot.2007.324.

Doorbar, John (2005): The papillomavirus life cycle. In *Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology* 32 Suppl 1, pp. S7-15. DOI: 10.1016/j.jcv.2004.12.006.

Doorbar, John (2006): Molecular biology of human papillomavirus infection and cervical cancer. In *Clinical science (London, England : 1979)* 110 (5), pp. 525–541. DOI: 10.1042/CS20050369.

Doorbar, John (2013): The E4 protein; structure, function and patterns of expression. In *Virology* 445 (1-2), pp. 80–98. DOI: 10.1016/j.virol.2013.07.008.

Fischer, Nina M.; Favrot, Claude; Birkmann, Katharina; Jackson, Michele; Schwarzwald, Colin C.; Müller, Martin et al. (2014): Serum antibodies and DNA indicate a high prevalence of equine papillomavirus 2 (EcPV2) among horses in Switzerland. In *Vet. Dermatol.* 25 (3), pp. 210-e54. DOI: 10.1111/vde.12129.

Franceschini, Andrea; Szklarczyk, Damian; Frankild, Sune; Kuhn, Michael; Simonovic, Milan; Roth, Alexander et al. (2013): STRING v9.1: protein-protein interaction networks, with increased coverage and integration. In *Nucleic acids research* 41 (Database issue), pp. D808-15. DOI: 10.1093/nar/gks1094.

Gérard, Claude; Goldbeter, Albert (2014): The balance between cell cycle arrest and cell proliferation: control by the extracellular matrix and by contact inhibition. In *Interface focus* 4 (3), p. 20130075. DOI: 10.1098/rsfs.2013.0075.

Ghim, Shin-Je; Rector, Annabel; Delius, Hajo; Sundberg, John P.; Jenson, A. Bennett; van Ranst, Marc (2004): Equine papillomavirus type 1: complete nucleotide sequence and characterization of recombinant virus-like particles composed of the EcPV-1 L1 major capsid protein. In *Biochem. Biophys. Res. Commun.* 324 (3), pp. 1108–1115. DOI: 10.1016/j.bbrc.2004.09.154.

Graham, S. V. (2010): Human papillomavirus: gene expression, regulation and prospects for novel diagnostic methods and antiviral therapies. In *Future Microbiol* 5 (10), pp. 1493–1506. DOI: 10.2217/fmb.10.107.

Haller, K.; Stubenrauch, F.; Pfister, H. (1995): Differentiation-dependent transcription of the epidermodysplasia verruciformis-associated human papillomavirus type 5 in benign lesions. In *Virology* 214 (1), pp. 245–255. DOI: 10.1006/viro.1995.0028.

Hebsgaard, S. M.; Korning, P. G.; Tolstrup, N.; Engelbrecht, J.; Rouzé, P.; Brunak, S. (1996): Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. In *Nucleic Acids Res* 24 (17), pp. 3439–3452.

Holloway, Amy; Storey, Alan (2014): A conserved C-terminal sequence of high-risk cutaneous beta-human papillomavirus E6 proteins alters localization and signalling of β1-integrin to promote cell migration. In *The Journal of general virology* 95 (Pt 1), pp. 123–134. DOI: 10.1099/vir.0.057695-0.

Howley PM, Lowy DR; Fields, Bernard N.; Knipe, David M. (1990): Fields Virology, 5th edition // Fields' virology. Papillomaviruses. 2. ed. New York, NY: Raven Press.

Hsiao, Yu-Ping; Yang, Jen-Hung; Wu, Wen-Jun; Lin, Meng-Hsuan; Sheu, Gwo-Tarng (2013): E6 and E7 of human papillomavirus type 18 and UVB irradiation corporately regulate interleukin-6 and interleukin-8 expressions in basal cell carcinoma. In *Experimental dermatology* 22 (10), pp. 672–674. DOI: 10.1111/exd.12223.

Huang, Da Wei; Sherman, Brad T.; Lempicki, Richard A. (2009a): Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. In *Nucleic acids research* 37 (1), pp. 1–13. DOI: 10.1093/nar/gkn923.

Huang, Da Wei; Sherman, Brad T.; Lempicki, Richard A. (2009b): Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. In *Nature protocols* 4 (1), pp. 44–57. DOI: 10.1038/nprot.2008.211.

Johansson, Cecilia; Schwartz, Stefan (2013): Regulation of human papillomavirus gene expression by splicing and polyadenylation. In *Nature reviews. Microbiology* 11 (4), pp. 239–251. DOI: 10.1038/nrmicro2984.

Kanehisa, M.; Goto, S. (2000): KEGG: kyoto encyclopedia of genes and genomes. In *Nucleic Acids Res* 28 (1), pp. 27–30.

Kanehisa, Minoru; Goto, Susumu; Sato, Yoko; Kawashima, Masayuki; Furumichi, Miho; Tanabe, Mao (2014): Data, information, knowledge and principle: back to metabolism in KEGG. In *Nucleic acids research* 42 (Database issue), pp. D199-205. DOI: 10.1093/nar/gkt1076.

Knight, C. G.; Munday, J. S.; Peters, J.; Dunowska, M. (2011): Equine penile squamous cell carcinomas are associated with the presence of equine papillomavirus type 2 DNA sequences. In *Vet. Pathol.* 48 (6), pp. 1190–1194. DOI: 10.1177/0300985810396516.

Knight, Cameron G.; Dunowska, Magda; Munday, John S.; Peters-Kennedy, Jeanine; Rosa, Brielle V. (2013): Comparison of the levels of Equus caballus papillomavirus type 2 (EcPV-2) DNA in equine squamous cell carcinomas and non-cancerous tissues using quantitative PCR. In *Vet. Microbiol.* 166 (1-2), pp. 257–262. DOI: 10.1016/j.vetmic.2013.06.004.

Lange, C. E.; Tobler, K.; Lehner, A.; Grest, P.; Welle, M. M.; Schwarzwald, C. C.; Favrot, C. (2013a): EcPV2 DNA in Equine Papillomas and In Situ and Invasive Squamous Cell Carcinomas Supports Papillomavirus Etiology. In *Vet. Pathol.* 50 (4), pp. 686–692. DOI: 10.1177/0300985812463403.

Lange, Christian E.; Tobler, Kurt; Ackermann, Mathias; Favrot, Claude (2011): Identification of two novel equine papillomavirus sequences suggests three genera in one cluster. In *Vet. Microbiol.* 149 (1-2), pp. 85–90. DOI: 10.1016/j.vetmic.2010.10.019.

Lange, Christian E.; Vetsch, Elisabeth; Ackermann, Mathias; Favrot, Claude; Tobler, Kurt (2013b): Four novel papillomavirus sequences support a broad diversity among equine papillomaviruses. In *J. Gen. Virol.* 94 (Pt 6), pp. 1365–1372. DOI: 10.1099/vir.0.052092-0.

Li, A.; Dubey, S.; Varney, M. L.; Dave, B. J.; Singh, R. K. (2003): IL-8 Directly Enhanced Endothelial Cell Survival, Proliferation, and Matrix Metalloproteinases Production and Regulated Angiogenesis. In *The Journal of Immunology* 170 (6), pp. 3369–3376. DOI: 10.4049/jimmunol.170.6.3369.

Li, Bo; Dewey, Colin N. (2011): RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. In *BMC Bioinformatics* 12, p. 323. DOI: 10.1186/1471-2105-12-323.

Lont, Anne P.; Kroon, Bin K.; Horenblas, Simon; Gallee, Maarten P W; Berkhof, Johannes; Meijer, Chris J L M; Snijders, Peter J F (2006): Presence of high-risk human papillomavirus DNA in penile carcinoma predicts favorable outcome in survival. In *Int J Cancer* 119 (5), pp. 1078–1081. DOI: 10.1002/ijc.21961.

Lu, Pengfei; Weaver, Valerie M.; Werb, Zena (2012): The extracellular matrix: a dynamic niche in cancer progression. In *The Journal of cell biology* 196 (4), pp. 395–406. DOI: 10.1083/jcb.201102147.

Lunardi, Michele; de Alcântara, Brígida Kussumoto; Otonel, Rodrigo Alejandro Arellano; Rodrigues, Wagner Borges; Alfieri, Alice Fernandes; Alfieri, Amauri Alcindo (2013): Bovine papillomavirus type 13 DNA in equine sarcoids. In *J. Clin. Microbiol.* 51 (7), pp. 2167–2171. DOI: 10.1128/JCM.00371-13.

Luo, Weijun; Brouwer, Cory (2013): Pathview: an R/Bioconductor package for pathway-based data integration and visualization. In *Bioinformatics (Oxford, England)* 29 (14), pp. 1830–1831. DOI: 10.1093/bioinformatics/btt285.

Mair, T. S.; Walmsley, J. P.; Phillips, T. J. (2000): Surgical treatment of 45 horses affected by squamous cell carcinoma of the penis and prepuce. In *Equine Veterinary Journal* 32 (5), pp. 406–410. DOI: 10.2746/042516400777591093.

Malumbres, Marcos; Barbacid, Mariano (2005): Mammalian cyclin-dependent kinases. In *Trends in biochemical sciences* 30 (11), pp. 630–641. DOI: 10.1016/j.tibs.2005.09.005.

McBride, Alison A. (2013): The papillomavirus E2 proteins. In *Virology* 445 (1-2), pp. 57–79. DOI: 10.1016/j.virol.2013.06.006.

Mine, Karina L.; Shulzhenko, Natalia; Yambartsev, Anatoly; Rochman, Mark; Sanson, Gerdine F O; Lando, Malin et al. (2013): Gene network reconstruction reveals cell cycle and antiviral genes as major drivers of cervical cancer. In *Nature communications* 4, p. 1806. DOI: 10.1038/ncomms2693.

Mootha, Vamsi K.; Lindgren, Cecilia M.; Eriksson, Karl-Fredrik; Subramanian, Aravind; Sihag, Smita; Lehar, Joseph et al. (2003): PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. In *Nature genetics* 34 (3), pp. 267–273. DOI: 10.1038/ng1180.

Mosseri, S.; Hetzel, U.; Hahn, Shelley; Michaloupoulou, Eleni; Sallabank, Hannah Clare; Knottenbelt, Derek C.; Kipar, A. (2014): Equine sarcoid: In situ demonstration

of matrix metalloproteinase expression. In *Veterinary journal (London, England : 1997)* 202 (2), pp. 279–285. DOI: 10.1016/j.tvjl.2014.07.026.

Muneer, Asif; Kayes, O.; Ahmed, Hashim U.; Arya, Manit; Minhas, Suks (2009): Molecular prognostic factors in penile cancer. In *World J Urol* 27 (2), pp. 161–167. DOI: 10.1007/s00345-008-0275-y.

Nagalakshmi, Ugrappa; Waern, Karl; Snyder, Michael (2010): RNA-Seq: a method for comprehensive transcriptome analysis. In *Curr Protoc Mol Biol* Chapter 4, pp. Unit 4.11.1-13. DOI: 10.1002/0471142727.mb0411s89.

Nigg, E. A. (1995): Cyclin-dependent protein kinases: key regulators of the eukaryotic cell cycle. In *BioEssays : news and reviews in molecular, cellular and developmental biology* 17 (6), pp. 471–480. DOI: 10.1002/bies.950170603.

Protzel, C.; Knoedel, J.; Zimmermann, U.; Woenckhaus, C.; Poetsch, M.; Giebel, J. (2007): Expression of proliferation marker Ki67 correlates to occurrence of metastasis and prognosis, histological subtypes and HPV DNA detection in penile carcinomas. In *Histol Histopathol* 22 (11), pp. 1197–1204.

Pruitt, Kim D.; Brown, Garth R.; Hiatt, Susan M.; Thibaud-Nissen, Francoise; Astashyn, Alexander; Ermolaeva, Olga et al. (2014): RefSeq: an update on mammalian reference sequences. In *Nucleic Acids Res* 42 (Database issue), pp. D756-63. DOI: 10.1093/nar/gkt1114.

R Core Team: R: A language and environment for statistical computing 2013. Available online at http://www.r-project.org./.

R Development Core Team (2013): R: A language and environment for statistical computing. Vienna, Austria.

Rajkumar, Thangarajan; Sabitha, Kesavan; Vijayalakshmi, Neelakantan; Shirley, Sundersingh; Bose, Mayil Vahanan; Gopal, Gopisetty; Selvaluxmy, Ganesharaja (2011): Identification and validation of genes involved in cervical tumourigenesis. In *BMC Cancer* 11, p. 80. DOI: 10.1186/1471-2407-11-80.

Regezi, Joseph A.; Dekker, Nusi P.; Ramos, Daniel M.; Li, Xiaowu; Macabeo-Ong, Maricris; Jordan, Richard C K (2002): Proliferation and invasion factors in HIV-associated dysplastic and nondysplastic oral warts and in oral squamous cell carcinoma: an immunohistochemical and RT-PCR evaluation. In *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* 94 (6), pp. 724–731. DOI: 10.1067/moe.2002.129760.

Robinson, James T.; Thorvaldsdóttir, Helga; Winckler, Wendy; Guttman, Mitchell; Lander, Eric S.; Getz, Gad; Mesirov, Jill P. (2011): Integrative genomics viewer. In *Nature biotechnology* 29 (1), pp. 24–26. DOI: 10.1038/nbt.1754.

Robinson, Mark D.; McCarthy, Davis J.; Smyth, Gordon K. (2010): edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. In *Bioinformatics* 26 (1), pp. 139–140. DOI: 10.1093/bioinformatics/btp616.

Robinson, Mark D.; Oshlack, Alicia (2010): A scaling normalization method for differential expression analysis of RNA-seq data. In *Genome Biol.* 11 (3), pp. R25. DOI: 10.1186/gb-2010-11-3-r25.

Ryzhakova, O. S.; Solov'eva, N. I. (2013): Matrix metalloproteinases (MMP)--MMP-1,-2,-9 and its endogenous activity regulators in transformed by E7 oncogene HPV16 and HPV18 cervical carcinoma cell lines. In *Biomed Khim* 59 (5), pp. 530–540.

Saito, Rintaro; Smoot, Michael E.; Ono, Keiichiro; Ruscheinski, Johannes; Wang, Peng-Liang; Lotia, Samad et al. (2012): A travel guide to Cytoscape plugins. In *Nature methods* 9 (11), pp. 1069–1076. DOI: 10.1038/nmeth.2212.

Santamaría, David; Barrière, Cédric; Cerqueira, Antonio; Hunt, Sarah; Tardy, Claudine; Newton, Kathryn et al. (2007): Cdk1 is sufficient to drive the mammalian cell cycle. In *Nature* 448 (7155), pp. 811–815. DOI: 10.1038/nature06046.

Santegoets, Lindy A M; van Baars, Romy; Terlou, Annelinde; Heijmans-Antonissen, Claudia; Swagemakers, Sigrid M A; van der Spek, Peter J et al. (2012): Different DNA damage and cell cycle checkpoint control in low- and high-risk human papillomavirus infections of the vulva. In *International journal of cancer. Journal international du cancer* 130 (12), pp. 2874–2885. DOI: 10.1002/ijc.26345.

Sarode, Gargi S.; Sarode, Sachin C. (2014): E6 oncoprotein interaction with paxillin and FAK. In *Oral oncology* 50 (4), pp. e17. DOI: 10.1016/j.oraloncology.2014.01.007.

Scase, T.; Brandt, S.; Kainzbauer, C.; Sykora, S.; Bijmholt, S.; Hughes, K. et al. (2010): Equus caballus papillomavirus-2 (EcPV-2): an infectious cause for equine genital cancer? In *Equine Vet. J.* 42 (8), pp. 738–745. DOI: 10.1111/j.2042-3306.2010.00311.x.

Schwartz, Stefan (2013): Papillomavirus transcripts and posttranscriptional regulation. In *Virology* 445 (1-2), pp. 187–196. DOI: 10.1016/j.virol.2013.04.034.

Scott, Danny W.; Miller, William H. (2011): Equine dermatology. 2nd ed. Maryland Heights, Mo: Elsevier/Saunders.

Shiau, Ming-Yuh; Fan, Li-Ching; Yang, Shun-Chun; Tsao, Chang-Hui; Lee, Huei; Cheng, Ya-Wen et al. (2013): Human papillomavirus up-regulates MMP-2 and MMP-9 expression and activity by inducing interleukin-8 in lung adenocarcinomas. In *PloS one* 8 (1), pp. e54423. DOI: 10.1371/journal.pone.0054423.

Subramanian, Aravind; Tamayo, Pablo; Mootha, Vamsi K.; Mukherjee, Sayan; Ebert, Benjamin L.; Gillette, Michael A. et al. (2005): Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. In *Proceedings of the National Academy of Sciences of the United States of America* 102 (43), pp. 15545–15550. DOI: 10.1073/pnas.0506580102.

Sykora, Sabine; Samek, Lisa; Schönthaler, Katharina; Palm, Franziska; Borzacchiello, Giuseppe; Aurich, Christine; Brandt, Sabine (2012): EcPV-2 is transcriptionally active in equine SCC but only rarely detectable in swabs and semen from healthy horses. In *Vet. Microbiol.* 158 (1-2), pp. 194–198. DOI: 10.1016/j.vetmic.2012.02.006.

Szklarczyk, Damian; Franceschini, Andrea; Kuhn, Michael; Simonovic, Milan; Roth, Alexander; Minguez, Pablo et al. (2011): The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. In *Nucleic acids research* 39 (Database issue), pp. D561-8. DOI: 10.1093/nar/gkq973.

Tan, Chye Ling; Gunaratne, Jayantha; Lai, Deborah; Carthagena, Laetitia; Wang, Qian; Xue, Yue Zhen et al. (2012): HPV-18 E2^E4 chimera: 2 new spliced transcripts and proteins induced by keratinocyte differentiation. In *Virology* 429 (1), pp. 47–56. DOI: 10.1016/j.virol.2012.03.023.

Thorvaldsdóttir, Helga; Robinson, James T.; Mesirov, Jill P. (2013): Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. In *Briefings in bioinformatics* 14 (2), pp. 178–192. DOI: 10.1093/bib/bbs017.

Tong, Xiao; Howley, Peter M. (1997): The bovine papillomavirus E6 oncoprotein interacts with paxillin and disrupts the actin cytoskeleton. In *Proc Natl Acad Sci U S A* 94 (9), pp. 4412–4417.

Van Den Top, J G B; Heer, N. de; Klein, W. R.; Ensink, J. M. (2008): Penile and preputial tumours in the horse: a retrospective study of 114 affected horses. In *Equine Vet. J.* 40 (6), pp. 528–532. DOI: 10.2746/042516408X281180.

Van den Top, J. G. B.; Harkema, L.; Lange, C.; Ensink, J. M.; van de Lest, C. H. A.; Barneveld, A. et al. (2014): Expression of p53, Ki67, EcPV2- and EcPV3 DNA, and viral genes in relation to metastasis and outcome in equine penile and preputial squamous cell carcinoma. In *Equine Vet J*, pp. n/a. DOI: 10.1111/evj.12245.

van Doorslaer, Koenraad; Tan, Qina; Xirasagar, Sandhya; Bandaru, Sandya; Gopalan, Vivek; Mohamoud, Yasmin et al. (2013): The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. In *Nucleic acids research* 41 (Database issue), pp. D571-8. DOI: 10.1093/nar/gks984.

Walker, Joanna; Smiley, Lucy Clare; Ingram, David; Roman, Ann (2011): Expression of Human Papillomavirus Type 16 E7 Is Sufficient to Significantly Increase Expression of Angiogenic Factors But Is Not Sufficient to Induce Endothelial Cell Migration. In *Virology* 410 (2), pp. 283–290. DOI: 10.1016/j.virol.2010.11.010.

Wang, Zhong; Gerstein, Mark; Snyder, Michael (2009): RNA-Seq: a revolutionary tool for transcriptomics. In *Nat. Rev. Genet.* 10 (1), pp. 57–63. DOI: 10.1038/nrg2484.

Waugh, David J J; Wilson, Catherine (2008): The interleukin-8 pathway in cancer. In *Clinical cancer research : an official journal of the American Association for Cancer Research* 14 (21), pp. 6735–6741. DOI: 10.1158/1078-0432.CCR-07-4843.

Westermarck, J.; Kahari, V. M. (1999): Regulation of matrix metalloproteinase expression in tumor invasion. In *FASEB J* 13 (8), pp. 781–792.

Yang, Bo-Chao; Wang, Feng-Xue; Zhang, Shu-Qin; Song, Ni; Li, Jian-Xi; Yang, Zhi-Qiang et al. (2013): Comparative evaluation of conventional polymerase chain reaction (PCR), with loop-mediated isothermal amplification and SYBR green I-based real-time PCR for the quantitation of porcine circovirus-1 DNA in contaminated samples destined for vaccine production. In *Journal of Virological Methods* 191 (1), pp. 1–8. DOI: 10.1016/j.jviromet.2013.03.014.

Yuan, Z. Q.; Nicolson, L.; Marchetti, B.; Gault, E. A.; Campo, M. S.; Nasir, L. (2008): Transcriptional changes induced by bovine papillomavirus type 1 in equine fibroblasts. In *J. Virol.* 82 (13), pp. 6481–6491. DOI: 10.1128/JVI.00429-08.

Yuan, ZhengQiang; Gobeil, Philipe A M; Campo, M. Saveria; Nasir, Lubna (2010): Equine sarcoid fibroblasts over-express matrix metalloproteinases and are invasive. In *Virology* 396 (1), pp. 143–151. DOI: 10.1016/j.virol.2009.10.010.

Zheng, Zhi-Ming; Baker, Carl C. (2006): Papillomavirus genome structure, expression, and post-transcriptional regulation. In *Front Biosci* 11, pp. 2286–2302.

Zinman, Guy E.; Naiman, Shoshana; Kanfi, Yariv; Cohen, Haim; Bar-Joseph, Ziv (2013): ExpressionBlast: mining large, unstructured expression databases. In *Nature methods* 10 (10), pp. 925–926. DOI: 10.1038/nmeth.2630.

Zur Hausen, Harald (2002): Papillomaviruses and cancer: from basic studies to clinical application. In *Nature reviews. Cancer* 2 (5), pp. 342–350. DOI: 10.1038/nrc798.

Acknowledgements

An erster Stelle muss ich meinem Betreuer Dr. Kurt Tobler danken, der mich in die Welt der Molekularbiologie und Datenanalyse eingeführt und in jeglicher Hinsicht über das ganze Projekt hinweg unterstützt hat. Er hatte immer ein offenes Ohr und hat so viel Input an Ideen gebracht, dass diese Dissertation nun doch etwas umfangreicher wurde als geplant. Zudem hat er es geschafft, mich so für dieses Feld zu motivieren, dass ich noch immer an der Virologie bin. Danke für alles!

Einen sehr großen Dank muss ich auch an Prof. Dr. Claude Favrot und Prof. Dr. Mathias Ackermann aussprechen, die mir die Möglichkeit gaben, diese Dissertation an der Universität Zürich zu verfassen. Zudem haben sie mich während des ganzen Projekts und beim Verfassen der Dissertation immer mit Anregungen begleitet.

Vielen Dank auch an Prof. Dr. Volker Thiel für die Verfassung des Korreferats.

Zudem großen Dank auch an das Functional Genomics Center Zürich, v.a. Catharine Aquino für die Durchführung der RNA-Seq und Unterstützung bei der Optimierung der RNA Qualität, sowie Lennart Oppitz für die Aufbereitung und ersten Analysen der Rohdaten.

Außerdem meinen herzlichen Dank an all die lieben Tierärzte die mich mit Proben versorgt haben, vor allem die Pferdekliniken der Vetsuisse Zürich und Bern, sowie Dr. Wolfgang Scheidemann, Dr. Bianca Schwarz und Dr. Maj Britt Cielewicz, aber auch Pferdemetzger Hans Rudolf Gloor.

Des Weiteren danke ich auch meinen Laborkollegen, vor allem Anita, Michi, Bruna, Cedi und Sereina die ich immer alles fragen konnte und meistens gute Ratschläge bekam. Zudem waren sie an langen Tagen immer eine gute Ablenkung zur Mittags- oder Kaffeepause und auch sonst, die half den Kopf wieder frei zu bekommen.

Vielen lieben Dank auch an meine Familie, vor allem an meine Mama Luise und meine Tante Ingrid. Ihr habt mich mein Leben lang unterstützt durch Schule, Studium und Doktorarbeit hinweg, bis jetzt zum nächsten Schritt dem Doktortitel. Danke dass ihr immer für mich da seid. Auch den anderen Familienmitgliedern, die nicht mehr auf dieser Erde weilen, gebührt mein herzlichster Dank für alle Unterstützung die ihr mir gabt und auch immer noch gebt.

Nicht unwesentlich beteiligt an dieser Arbeit war auch mein Pferd Caravaggio der mich nun seit 17 Jahren begleitet und die beste Abwechslung ist um den Kopf frei zu bekommen. Doch muss er sich aber auch bei langen Ausritten sämtliche Vorträge von mir anhören und auf seinem Rücken kommen mir immer wieder gute Ideen.

Zuletzt noch vielen Dank an all meine Freunde, v.a. Thorsten, Johanna, Saskia, Maxi und Susi, die keine so genaue Ahnung davon haben was ich da genau mache, aber sich dennoch immer interessiert meine Probleme anhören und mich wieder aufbauen, wenn es mal nicht so läuft wie geplant.

Supplementary Tables

Supplementary Table 1: affected genes in KEGG pathways

hgnc_symbol	log2 Ratio	p-value	fdr	hsa03030	hsa04110	hsa04512	hsa04510
BAD	3,109	0,00001852	0,0008056				X
BUB1	2,221	0,00057	0,00996		X		
BUB1B	1,955	0,002056	0,02411		X		
CAPN2	2,064	9,703E-06	0,0005194				X
CAV2	2,178	0,00006066	0,002015				X
CCNA2	2,467	0,0005901	0,01021		X		
CCNB1	3,192	0,00003809	0,001414		X		
CCNB2	2,103	0,005815	0,04933		X		
CCNB3	3,236	0,0002414	0,005431		X		
CCND2	2,879	0,0001073	0,00308		X		X
CDC14A	-2,549	0,0001073	0,00308		X		
CDC20	2,609	0,0003003	0,006375		X		
CDC25B	1,663	0,006335	0,05213		X		
CDC25C	2,632	0,0005934	0,01023		X		
CDC6	2,651	0,0002071	0,004873		X		
CDK1	3,424	0,00006274	0,002059		X		
CDK4	1,811	0,0007708	0,0122		X		
COL11A1	5,407	0,00001263	0,0006182			X	X
COL1A1	2,412	0,0006881	0,01129			X	X
COL3A1	2,064	0,0007525	0,01203			X	X
COL4A1	2,909	1,445E-08	2,872E-06			X	X
COL4A2	2,014	0,0001794	0,004409			X	X
COL5A1	1,676	0,008817	0,06499			X	X
COL5A2	1,871	0,0002378	0,005366			X	X
COMP	4,163	0,002266	0,02594			X	X
CTNNB1	1,419	0,001338	0,01811				X
DBF4	1,546	0,004163	0,03958		X		
DOCK1	-1,286	0,005054	0,04505				X
E2F1	1,899	0,007072	0,05603		X		
FEN1	1,974	0,002205	0,02546	X			
FLNC	-2,231	0,00789	0,06002				X
FN1	1,736	0,009121	0,06626			X	X
GADD45A	3,132	0,00006793	0,002193		X		
HDAC1	1,421	0,001223	0,017		X		
HMMR	2,823	0,0002259	0,005172			X	
IBSP	6,451	0,006076	0,05075			X	X
ITGA3	3,267	0,00001596	0,0007248			X	X
ITGA6	1,4	0,008375	0,06246			X	X
ITGB1	2,577	7,569E-06	0,0004308			X	X
ITGB4	1,755	0,002538	0,02818			X	X
JUN	1,754	0,0004201	0,008004				X

LAMB3	1,581	0,009048	0,06604			X	X
LAMC2	1,83	0,008117	0,06114			X	X
LIG1	1,327	0,007152	0,05648	X			
MAD2L1	1,31	0,009139	0,06626		X		
MCM2	1,805	0,002316	0,02641	X	X		
MCM3	1,747	0,001537	0,01974	X	X		
MCM4	2,047	0,001029	0,01493	X	X		
MCM5	2,298	0,0005877	0,01018	X	X		
MCM6	2,075	0,0006058	0,01037	X	X		
MCM7	1,946	0,0006018	0,01032	X	X		
MET	1,327	0,004514	0,04162				X
MYL12A	1,689	0,0002891	0,006187				X
PAK1	2,599	0,0002107	0,004923				X
PAK7	-6,942	1,646E-15	1,613E-12				X
PARVA	2,342	0,001226	0,017				X
PCNA	1,796	0,001804	0,02206	X	X		
PDGFC	1,503	0,00841	0,06266				X
PDPK1	-1,459	0,001349	0,01818				X
PIK3CD	-2,307	3,022E-06	0,0002115				X
PLK1	2,914	0,0001397	0,003726		X		
POLA2	2,118	0,0008051	0,01259	X			
POLE4	2,037	0,0005572	0,00988	X			
PTTG1	-1,286	0,005054	0,04505		X		
RAP1B	2,139	0,002041	0,02399				X
RFC2	1,619	0,006321	0,05208	X			
RFC3	1,78	0,005165	0,04551	X			
RFC4	2,367	0,0004178	0,00799	X			
RNASEH2B	1,727	0,002608	0,02869	X			
RNASEH2C	2,891	0,003686	0,03619	X			
RPA3	1,892	0,0003866	0,007598	X			
SDC4	1,581	0,001354	0,0182			X	
SFN	1,41	0,002577	0,02848		X		
SHC4	-3,036	0,003904	0,03771				X
SMAD2	1,304	0,009543	0,06798		X		
SMC1B	7,146	7,343E-07	0,00006995		X		
SPP1	4,185	0,0001543	0,00397			X	X
SSBP1	1,673	0,001628	0,02059	X			
THBS1	2,066	0,00072	0,01163			X	X
THBS2	3,791	1,916E-06	0,0001452			X	X
THBS4	4,354	0,00001266	0,0006182			X	X
TNC	3,637	1,648E-07	0,00002113			X	X
TNXB	-4,07	0,00006213	0,002051			X	X
TTK	2,164	0,00104	0,01503		X		
VASP	1,717	0,0009129	0,01385				X
VAV1	1,871	0,002142	0,02483				X
VAV2	1,89	0,007422	0,05768				X

Supplementary Table 2: 100 most upregulated genes depending on the p-value

hgnc_symbol	log2 ratio	p-value	fdr
PTHLH	6,85	2,69E-11	1,37E-08
ATP8A2	11,05	1,22E-10	5,40E-08
SLPI	7,04	1,86E-10	7,49E-08
S100A6	3,62	2,01E-10	7,87E-08
ALDH1A2	5,13	3,33E-10	1,20E-07
SDK2	5,42	3,75E-10	1,32E-07
PHLDA1	5,29	3,96E-10	1,36E-07
ISG15	4,51	7,39E-10	2,42E-07
SLC39A4	6,65	1,18E-09	3,70E-07
ID1	3,20	1,19E-09	3,70E-07
IL8	**13,85**	**1,36E-09**	**4,03E-07**
MMP13	9,58	1,43E-09	4,09E-07
LTBP1	2,90	1,85E-09	5,08E-07
TSPAN15	4,85	2,30E-09	6,06E-07
IFIT1	3,75	4,21E-09	1,01E-06
C14orf164	4,33	6,10E-09	1,39E-06
FHL2	5,16	7,66E-09	1,72E-06
PITX1	6,61	1,15E-08	2,43E-06
COL4A1	2,91	1,45E-08	2,87E-06
S100A11	3,23	1,72E-08	3,28E-06
MMP1	**9,43**	**2,00E-08**	**3,76E-06**
SLC16A3	3,77	2,31E-08	4,28E-06
TREM1	12,26	3,23E-08	5,68E-06
SLC9B2	3,57	3,58E-08	6,15E-06
TMPRSS4	4,41	3,59E-08	6,15E-06
S100A2	5,32	4,62E-08	7,64E-06
RBP1	3,79	5,23E-08	8,45E-06
MYBL2	4,34	6,05E-08	9,54E-06
MYO5A	2,92	6,15E-08	9,58E-06
MMP9	7,19	6,34E-08	9,77E-06
CCL2	5,00	6,86E-08	1,04E-05
EPOR	5,76	6,99E-08	1,04E-05
IGF2BP2	6,30	7,78E-08	1,15E-05
KRT14	3,18	8,92E-08	1,29E-05
ACOT9	3,12	9,19E-08	1,30E-05
NDRG1	3,66	9,38E-08	1,31E-05
AGR2	8,06	1,13E-07	1,54E-05
RND1	3,48	1,40E-07	1,85E-05
SATL1	8,16	1,59E-07	2,06E-05
TNC	3,64	1,65E-07	2,11E-05
BACE2	2,98	1,83E-07	2,32E-05
RGL3	5,83	2,60E-07	3,16E-05
PLAUR	5,46	3,42E-07	4,01E-05
PMEPA1	3,93	3,47E-07	4,03E-05
BNC1	3,47	3,69E-07	4,22E-05
MSANTD3	3,77	3,89E-07	4,41E-05
CWH43	5,03	4,23E-07	4,68E-05
SULF1	3,96	4,45E-07	4,85E-05
SERPINE2	3,87	4,58E-07	4,95E-05
S100A10	3,21	4,79E-07	5,09E-05

hgnc_symbol	log2 ratio	p-value	fdr
CXCL2	5,45	5,83E-07	5,99E-05
FRMPD4	5,04	5,83E-07	5,99E-05
IRF7	3,46	6,40E-07	6,45E-05
SLC1A5	2,88	6,64E-07	6,60E-05
PTX3	5,94	6,96E-07	6,82E-05
CSF3R	5,66	7,03E-07	6,84E-05
SMC1B	7,15	7,34E-07	7,00E-05
C9orf89	3,52	7,45E-07	7,05E-05
STK31	11,10	7,51E-07	7,06E-05
CA9	4,61	8,61E-07	7,95E-05
PPA1	2,95	8,64E-07	7,95E-05
LGALS9C	3,02	9,32E-07	8,47E-05
TXNDC17	2,41	9,82E-07	8,75E-05
TXN	3,11	9,89E-07	8,75E-05
MUC20	3,18	9,99E-07	8,79E-05
SAT1	2,83	1,05E-06	9,08E-05
CD68	3,23	1,08E-06	9,30E-05
GLIPR1	2,40	1,27E-06	1,08E-04
KDELR3	4,05	1,28E-06	1,08E-04
TNS4	3,65	1,34E-06	1,10E-04
FGD2	3,87	1,38E-06	1,13E-04
ARSJ	2,90	1,50E-06	1,20E-04
FCGBP	2,58	1,60E-06	1,27E-04
SYCE2	7,01	1,84E-06	1,42E-04
PDLIM7	2,68	1,90E-06	1,44E-04
THBS2	3,79	1,92E-06	1,45E-04
CCL4	3,03	2,12E-06	1,56E-04
PIWIL1	10,39	2,29E-06	1,68E-04
CXCL6	7,68	2,32E-06	1,69E-04
ARG2	3,22	2,82E-06	2,02E-04
CYP27B1	7,21	2,85E-06	2,02E-04
GPR110	3,81	2,91E-06	2,05E-04
SERPINB6	2,70	3,13E-06	2,18E-04
PADI3	7,34	3,16E-06	2,19E-04
DZIP1	2,40	3,27E-06	2,25E-04
IRG1	6,80	3,51E-06	2,40E-04
UBE2C	3,70	3,77E-06	2,50E-04
ALDH3A1	4,70	3,82E-06	2,52E-04
KCNE3	4,33	4,41E-06	2,87E-04
IGFBP2	2,33	4,66E-06	2,97E-04
ZFYVE28	8,37	4,67E-06	2,97E-04
ANXA1	3,41	4,70E-06	2,97E-04
ADAMDEC1	9,67	4,80E-06	3,01E-04
MX2	4,33	5,11E-06	3,15E-04
GPX8	2,96	5,22E-06	3,21E-04
ZNF114	3,71	5,69E-06	3,48E-04
P4HA3	3,75	6,35E-06	3,77E-04
DSG2	2,96	6,64E-06	3,91E-04
MARCKSL1	2,91	6,83E-06	3,99E-04
SPARC	2,50	7,37E-06	4,25E-04

Supplementary Table 3: "EXPRESSION BLAST" matches no preselection

Nr	GSE	PMID	Description
0		This study	Top 100 of this study (Supplementary Table2)
1	12630	PMID:19332734	Gene expression profiles of poorly differentiated undifferentiated and metastatic cancers
2	33116		Breast cancer and liver tissue biopsies used to develop and validate an index for liver contamination in metastatic breast cancer
3	20916	PMID:20957034	Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data
4	27157	PMID:22276141	Prognostic Significance and Gene Expression Profiles of p53 Mutations in Microsatellite Stable Stage III Colorectal Adenocarcinoma
5	22563		Identification of TSGs Frequently Methylated in Renal Cell carcinoma expression profiles of Renal cell lines following de methylation
6	43029	PMID:23630276	RNASET2 silenced and control OVCAR3 cell clones injected into nude mice
7	4271	PMID:16530701	Molecular subclasses of high grade glioma prognosis disease progression and neurogenesis
8	29721	PMID:21747116	The landscape of promoter DNA hypomethylation in liver cancer expression data
9	25251	PMID:21684623	Establishment and Comparative Characterization of Novel Non Small Cell Lung Cancer Cell Lines and Their Corresponding Tumor Tissue
10	21122	PMID:20601955	Whole transcript expression data for soft tissue sarcoma tumors and control normal fat specimens
11	22544	PMID:20799942	Integration of transcript expression copy number and LOH analysis of infiltrating ductal carcinoma of the breast expression analysis
12	27480	PMID:21952923	Gene expression analysis of Oncostatin M OSM signalling in cervical squamous cell carcinomas over expressing the Oncostatin M receptor OSMR
13	9750	PMID:18506748	Identification of gene expression profiles in cervical cancer

14	21687	PMID:20639864	Comparative genomics matches mutations and cells to generate faithful ependymoma models
15	9844	PMID:18254958	Transcriptomic Dissection of Tongue Squamous Cell Carcinoma
16	41258	PMID:19359472	Expression data from colorectal cancer patients
17	44408		Transcriptomic survey of lymph node positive vs negative ductal breast cancer
18	10972	PMID:18538736	Colon cancer
19	5364	PMID:18636107	A Precisely Regulated Gene Expression Cassette Potently Modulates Metastasis and Survival in Multiple Solid Cancers
20	29570		The mtDNA Amerindian Haplogroup B2 enhances the risk for Cervical Cancer of HPV de regulation of mitochondrial genes may be involved
21	6956	PMID:18245496	Tumor Immunobiological Differences in Prostate Cancer between African American and European American Men
22	17674	PMID:22084725	Inflammatory gene profiling of Ewing sarcoma family of tumors set B
23	36895	PMID:22683710	Molecular Genetic Classification of clear cell Renal Cell Carcinoma ccRCC based on the Gene Expression Profiling of Tumors and Tumorgrafts deficient for BAP1 or PBRM1
24	27854	PMID:23065711	Overexpression of NUCKS1 in colorectal cancer correlates with recurrence after curative surgery gene expression analysis
25	17679	PMID:22084725	Inflammatory gene profiling of Ewing sarcoma family of tumors

Supplementary Table 4: "EXPRESSION BLAST" matches preselection "carcinoma"

Nr	GSE	PMID	Description
0		This study	Top 100 of this study (Supplementary Table 2)
1	13601	PMID:19138406	Oral tongue cancer gene expression profiling Identification of novel potential prognosticators
2	20916	PMID:20957034	Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data
3	22563		Identification of TSGs Frequently Methylated in Renal Cell carcinoma expression profiles of Renal cell lines following de methylation
4	29721	PMID:21747116	The landscape of promoter DNA hypomethylation in liver cancer expression data
5	6008	PMID:16452189	Human ovarian tumors and normal ovaries
6	22544	PMID:20799942	Integration of transcript expression copy number and LOH analysis of infiltrating ductal carcinoma of the breast expression analysis
7	18462		Comparison of gene expression profiles between paired primary and metastasis colorectal carcinoma
8	27155	PMID:16007166	Human thyroid adenomas carcinomas and normals
9	9844	PMID:18254958	Transcriptomic Dissection of Tongue Squamous Cell Carcinoma
10	41258	PMID:19359472	Expression data from colorectal cancer patients
11	21422	PMID:21314937	Expression profiling of human DCIS and invasive ductal breast carcinoma
12	36895	PMID:22683710	Molecular Genetic Classification of clear cell Renal Cell Carcinoma ccRCC based on the Gene Expression Profiling of Tumors and Tumorgrafts deficient for BAP1 or PBRM1
13	14323	PMID:19098997	RMA expression data for liver samples from subjects with HCV HCV HCC or normal liver
14	33426	PMID:22280838	Three Dimensional Tumor Profiling Reveals Minimal mRNA Heterogeneity in Esophageal Squamous Cell Carcinoma
15	39612	PMID:23223137	Distinct gene expression profiles of viral and non viral associated Merkel cell carcinoma revealed by transcriptome analysis
16	17856	PMID:20380719	Gene expression in nontumoral liver tissue and recurrence free survival in hepatitis C virus positive HCC

17	16873	PMID:19700746	Early Dysregulation of Cell Adhesion and Extracellular Matrix Pathways in Breast Cancer Progression
18	23772		Why does HAMLET preferentially kill tumor cells p38 dependent death in tumor but up regulation of innate immunity in healthy differentiated cells
19	9115	PMID:17981789	Determination of a cell proliferation and chromosomal instability signature in anaplastic thyroid carcinoma
20	11260	PMID:19212673	Expression profiling of HCC patients with different recurrent free survival time
21	6465	PMID:18490075	Expression data of Hepatocellular Carcinoma
22	4183	PMID:19461970	Inflammation adenoma and cancer objective classification of colon biopsy specimens with gene expression signature
23	42109		Identification of anaplastic lymphoma kinase as a candidate of new therapeutic target for BCC
24	14773	PMID:21640118	Roles of EMT regulator in colon cancer
25	45645		cDNA microarray data after ATP treatment